VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

DEPARTMENT OF BIOMEDICAL ENGINEERING

SOFTWARE PRO EXTRAKCI GENOMICKÝCH DAT Z GENBANK FORMÁTU

USER INTERFACE FOR DATA EXTRACTION FROM GENBANK DATA FORMAT

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

AUTOR PRÁCE — KATEŘINA JURE ČKOVÁ

AUTHOR

VEDOUCÍ PRÁCE — Ing. DENISA MADĚRÁNKOVÁ

SUPERVISOR

BRNO 2015

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

Fakulta elektrotechniky a komunikačních technologií

Ústav biomedicínského inženýrství

Bakalářská práce

bakalářský studijní obor
Biomedicínská technika a bioinformatika

Studentka: Kateřina Jurečková ***ID:*** 154636
Ročník: 3 ***Akademický rok:*** 2014/2015

NÁZEV TÉMATU:

Software pro extrakci genomických dat z GenBank formátu

POKYNY PRO VYPRACOVÁNÍ:

1) Seznamte se s databází GenBank a používaným formátem dat gbk, formát podrobně popište a srovnejte s dalšími používanými formáty sekvenčních dat. 2) Seznamte se se strukturou mitochondriální DNA živočichů a rostlin, u rostlin dále také plastidovou DNA. 3) Vyhodnoťte úspěšnost automatické extrakce dat z gbk formátu pro celé mitochondriální sekvence funkcí genbankread z Matlabu. 4) V programovém prostředí Matlab vytvořte grafickou uživatelskou aplikaci pro extrakci dat dle jednotlivých genů mitochondriální a plazmidové DNA s různým nastavením parametrů výstupního formátu. 5) Pomocí aplikace sestavte datasety sekvencí jednotlivých genů mitochondriální a plastidové DNA různých organismů a vyhodnoťte vnitrodruhové a mezidruhové variability. 6) Výsledky práce diskutujte.

DOPORUČENÁ LITERATURA:

[1] SCHEFFLER, I. E. Mitochondria. 2nd ed., Wiley-Blackwell, 2007, 472 s. ISBN 978-0-470-04073-7.
[2] BENDICH, A. Circular Chloroplast Chromosomes: The Grand Illusion. The Plant Cell. 2004, 16, 1661-1666.

Termín zadání: 9. 2. 2015 ***Termín odevzdání:*** 29. 5. 2015

Vedoucí práce: Ing. Denisa Maděránková
Konzultanti semestrální práce:

prof. Ing. Ivo Provazník, Ph.D.
Předseda oborové rady

UPOZORNĚNÍ:

Abstrakt:
Tato bakalářská práce obsahuje literární rešerši na témata databáze GenBank, její nejpoužívanější datový formát gbk a jeho srovnání s dalšími používanými formáty sekvenčních dat. Dále práce popisuje strukturu mitochondriální DNA živočichů a rostlin, u rostlin blíže popisuje i DNA plastidovou. V praktické části se práce zaměřuje na vyhodnocení úspěšnosti automatické extrakce dat z gbk formátu pro celé mitochondriální sekvence pomocí funkce `genbankread` z Matlabu. A popisuje nově vytvořenou aplikaci na extrakci dat z GenBank souborů. V závěru je tato aplikace využita při analýze mitochondriálních genomů.

Klíčová slova:
Databáze GenBank, mitochondriální DNA, genbankread, rodová a mezidruhová variabilita

Abstract:
This bachelor thesis contains a literature review on the topic of the GenBank database, its most used flat file format gbk and its comparison with other formats of sequential data. Furthermore this thesis describes the structure of mitochondrial DNA of animals and plants and plastid DNA of plants. In the practical part of this thesis is evaluation of successful automatic data extraction from GenBank flat file format from whole mitochondrial sequences by `genbankread` function in Matlab. And it describes also a new application for GenBank data extraction. In the end this application is used in analyses of mitochondrial genomes.

Keywords:
GenBank database, mitochondrial DNA, genbankread, genus and interspecific variability

Bibliografická citace:

JUREČKOVÁ, K. *Software pro extrakci genomických dat z GenBank formátu.* Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2015. 50 s. Vedoucí semestrální práce Ing. Denisa Maděránková.

Prohlášení

Prohlašuji, že svoji bakalářské práci na téma „Software pro extrakci genomických dat z GenBank formátu“ jsem vypracovala samostatně pod vedením vedoucího bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autorka uvedené bakalářské práce dále prohlašuji, že v souvislosti s vytvořením této práce jsem neporušila autorská práva třetích osob, zejména jsem nezasáhla nedovoleným způsobem do cizích autorských práv osobnostních a jsem si plně vědoma následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení § 152 trestního zákona č. 140/1961 Sb.

V Brně dne 29. května 2015 ……

podpis autorky

Poděkování

Děkuji vedoucí bakalářské práce Ing. Denise Maděránkové za účinnou metodickou, pedagogickou a odbornou pomoc a další cenné rady při zpracování mé bakalářské práce.

V Brně dne 29. května 2015 ……

podpis autorky

Obsah

Seznam obrázků

Seznam tabulek

Úvod

Tato bakalářská práce se zaměřuje na problematiku využití dat získaných z veřejně přístupných databází různých biologických sekvencí. Zejména se soustřeďuje na databázi GenBank, která obsahuje všechny veřejně dostupné nukleotidové sekvence včetně bližších informací. Nejčastěji používaný datový formát se jmenuje GenBank a blíže je popsán v této práci v kapitole číslo 1. Stejně tak i datový formát FASTA, který redukuje komplexní informace z GenBank formátu pouze na základní údaje.

Dalším tématem této práce je DNA mitochondriální a plastidová, jejichž sekvence lze nalézt v GenBank databázi. Mitochondriální DNA je stejně jako plastidová DNA nositelkou mimojaderné dědičnosti. Mitochondriální DNA lze nalézt v matrix mitochondrie, což je jedna z nejvýznamnějších organel živočišné i rostlinné buňky. Plastidová DNA se nachází v chloroplastech, a proto se někdy též říká DNA chloroplastová, a nalezneme ji pouze u rostlin. Struktura těchto DNA je popsána v kapitole číslo 2.

V praktické části jsou nejdříve analyzovány možnosti zpracování záznamů z GenBank databáze pomocí programového prostředí Matlab a bioinformatického toolboxu. Zejména je prověřena možnost načtení GenBank souborů pomocí funkce `genbankread` a jsou prověřeny její možnosti. Na základě této analýzy poté byla vytvořena uživatelská aplikace, která dokáže tuto funkci nahradit a nabídnout více možností pro práci s GenBank souborem. Blíže je tato aplikace popsána v kapitole číslo 4.2.

V neposlední řadě byla aplikace využita k vytvoření datasetů mitochondriálních sekvencí a tyto datasety byly podrobeny rodové a mezidruhové analýze. Výsledky těchto analýz je možné nalézt v kapitole číslo 5.

1. Databáze GenBank

GenBank je databáze obsahující veškeré veřejně přístupné nukleotidové sekvence. Příspěvky databáze GenBank mohou být také sekvence mRNA s kódujícími úseky, úseky genomické DNA pro jeden nebo více genů anebo celé genomy organismů. Tyto příspěvky jsou dále děleny do skupin podle fylogenetického původu nebo podle metody, díky které byly tyto informace získány. Celý obsah databáze tvoří příspěvky přímo od samotných autorů, kteří jsou ochotni se o výsledky své práce podělit. Tato databáze a jí podobné tvoří v dnešní době nepostradatelnou součást molekulární biologie. Díky nim lze předávat informace o výsledcích výzkumů z celého světa pomocí počítače a připojení k internetu [12].

GenBank spadá pod the National Institutes of Health (NIH) a provozuje ji the Nation Center for Biotechnology Information (NCBI, sídlící v Bethsedě v Marylandu v USA). GenBank patří do skupiny mezinárodních databází, které navzájem sdílí svůj obsah. Do této skupiny patří dále the DNA Data Bank of Japan (DDBJ, sídlící v Mishimě v Japonsku) a také the European Molecular Biology Laboratory (EMBL), která patří pod the European Bioinformatics Institute (EBI, sídlící v Hinxtonu ve Velké Británii). V principu není důležité, do které ze tří výše uvedených databází je příspěvek vložen, protože výměna informací a synchronizace mezi nimi probíhá každý den. Výsledkem je tedy jednotná databáze[3].

Mimojiné NCBI spravuje ještě jednu databázi sekvencí – RefSeq. Tato databáze je na rozdíl od GenBank dozorovaná a vytváří ji pouze NCBI a to z existujících záznamů v GenBank. NCBI se dále stará o tyto příspěvky a případně je upravuje podle nových dat. Také na rozdíl od GenBanku lze nalézt pro jeden zajímavý úsek právě jeden záznam. RefSeq vytváří tyto zkontrolované záznamy pouze pro omezený počet druhů.

1.1. Historie GenBank

Když vědci poprvé začali analyzovat proteiny a DNA byl to velmi nákladný a zdlouhavý proces. Navíc byl často limitován pouze na určité geny nebo proteiny, které vědce zajímali. Malé skupinky vědců začali tyto data shromažďovat a porovnávat. Tak náhodně zjišťovali, že například některé dva proteiny jsou evolučně příbuzné.

Na konci sedmdesátých let dvacátého století bylo jasné, že je potřeba vytvořit mezinárodní počítačovou databázi pro nukleotidové sekvence. Proto NIH pořádal od roku 1979 sérii workshopů, během kterých vznikali projekty pro tuto novou databázi. V roce 1982 NIH určil jako vítězný projekt práci Bolta, Beraneka a Newmana z Los Alamos National Laboratory, kteří obdrželi pětiletý kontrakt na databázi nukleových sekvencí. A tím byl položen základ databáze. GenBank ovšem nebyl první databází svého druhu na světě, první byla databáze EMBL ve Velké Británii. Nicméně po

několika letech již spolu obě databáze spolupracovali a nakonec i s příchodem DDBJ vznikla v polovině osmdesátých let the International Nucleotide Sequence Database Collaboration (INSDC). Následně byl růst databází podmíněn vědeckými časopisy, které začali vyžadovat od autorů přístupová čísla k GenBank, EMBL nebo DDBJ u článků, které se týkaly sekvencí nukleotidů.

V roce 1987 NIH uzavřel kontrakt s firmou IntelliGenetics opět pod Los Alamos National Laboratory a v roce 1992 byla databáze GenBank přesunuta do NCBI.

V dnešní době je GenBank propojen s mnoha dalšími biologickými databázemi (genomické mapy, proteinové struktury), tak i s vědeckými články (PubMed Central Database) či nástroji pro analýzu dat. Vzhledem k obrovskému pokroku v oblasti sekvenovacích technologií je získáváno obrovské množství dat každý den a velikost objemu dat v databázi roste exponenciálně [4], [5].

1.2. Používané datové formáty v databázi GenBank

Databáze obsahují příspěvky ve formě sekvencí a užitečné anotace. Výstupem databáze můžou být různé datové formáty například GenBank Flatfile, FASTA a další.

1.2.1. GenBank

Datový formát GenBank je základem celé stejnojmenné databáze. Jedná se o nejrozšířenější a nejpoužívanější formát pro vyjádření informací o nukleotidových sekvencích. Jelikož se informace neustále doplňují do všech tří databází, musí být možné jednotlivé formáty mezi sebou převádět. Co se týče formátu používaného v databázi DDBJ, tak se jedná o téměř identický s GenBank. Oproti tomu EMBL používá na začátku každého řádku speciální předponu a ta určuje, jaký druh informace je v daném řádku uveden [3].

Datový formát GenBank se skládá ze tří částí. První část je hlavička, kde jsou uvedeny hlavní informace o daném příspěvku. Další částí je anotace, která je zde podrobně rozepsána do jednotlivých sekcí, jak bude uvedeno níže. A poslední součástí je již samotná nukleotidová sekvence [7], [8].

Hlavička (Locus)

```
LOCUS   NC_000005    3824 bp   DNA     linear    CON 03-FEB-2014
```

Hlavička obsahuje informace o jméně, délce sekvence, typu a tvaru molekuly, skupině, do které v rámci databáze spadá, a o datu poslední úpravy.

Název (Locus Name)

V ukázce se jedná o kód NC_000005 obsahující kombinaci velkých písmen a čísel většinou s maximální délkou deseti znaků. Původně bylo jméno vymyšleno tak, aby

reprezentovalo předmět daného záznamu a zároveň byl mnemotechnický. Dříve měl tento kód sloužit zařazení příspěvků se stejnými sekvencemi do stejné skupiny, dnes to už ovšem neplatí. V případě, že číselný kód má pět znaků, tak předcházející písmena značí rod a druhové jméno, číslo samotné je číslem přístupovým. Pokud je znaků osm (dvě písmena následována šesti číslicemi), samotný název je i přístupovým kódem. Pro vyhledávání v databázi je lepší používat přístupové číslo než název (locus name), protože tyto čísla jsou stálá a jedinečná, kdežto názvy se mohou měnit.

Délka sekvence (Sequence Length)
Toto číslo udává kolik párů bází (pb) je obsaženo v daném záznamu. Při vkládání sekvence do databáze je počet bází omezen. Výjimkou je informace o celém genomu, pro ten vytvoří databáze pouze jeden souvislý záznam. V ostatních případech je maximální limit omezen na 350 000 pb a spodní hranice je stanovena na 50 pb.

Typ molekuly (Molecule Type)
Informuje, o jaký typ molekuly se v záznamu jedná. V databázi lze vyhledat následující typy molekul:

- genomická DNA/RNA
- mRNA – mediátorová RNA
- rRNA – ribozomální RNA
- cRNA – komplementární RNA
- scRNA – malá cytoplazmická RNA
- snRNA – malá jaderná RNA
- snoRNA – malá jadérková RNA
- tRNA – transferová RNA

Topologie molekuly (Molecule Topology)
Topologie molekuly říká, jaký tvar má příslušná sekvence. Většinou se jedná o tvar lineární nebo u mitochondriální a plastidové DNA se jedná o tvar kruhový.

GenBank divize (GenBank Division)
V rámci databáze jsou sekvence rozděleny do níže uvedených skupin. Kritéria pro rozdělení jsou různá. Dříve byly sekvence rozdělovány do specifických skupin organismů, teď se rozdělují i podle specifických způsobů získání sekvence.

- PRI – primáti
- ROD – hlodavci

- MAM – ostatní savci
- VRT – další obratlovci
- INV – bezobratlovci
- PLN – rostliny, plísně a řasy
- BCT – bakterie
- VRL – viry
- PHG – bakteriofágové
- CON – sekvence takto označené, jsou pouze vyčleněné části primárních sekvencí
- SYN – syntetické sekvence
- UNA – neokomentovaná sekvence
- EST – expressed sequence tag
- PAT – patentované sekvence
- STS – sequence tagged site
- GSS – genome survey sequences
- HTG – high-throughput genomic sekvence
- HTC – nedokončená HTG sekvence
- ENV – sekvence z neznámého organismu

Datum změny (Modification Date)
Posledním údajem v hlavičce informuje o datu poslední úpravy příspěvku.

Definice (Definition)

```
DEFINITION  Homo sapiens chromosome 5, GRCh38 Primary Assembly.
```

Shrnuje nejdůležitější informace o sekvenci. Definice zahrnuje informaci o organismu, ze kterého sekvence pochází, případně jméno genu, proteinu, chromozomu anebo krátký popis funkce.

Přístupové číslo (Accession)

```
ACCESSION  NC_000005 REGION: 138465492..138469315  GPC_000001297
```

Jedná se o unikátní přístupový kód, který jednoznačně identifikuje záznam o sekvenci. Přístupové číslo je uděleno kompletnímu příspěvku a je to obvykle kombinace písmen a

čísel. Například to může být kombinace jednoho písmene a pěti čísel (U12345) nebo dvou písmen a šesti čísli (AF123456). Některé kódy mohou být delší v závislosti na typu záznamu. Přístupová čísla se záznamů se nikdy nemění a to ani v případě, že jsou informace v záznamu upraveny samotným autorem.

Záznamy z RefSeq databáze mají jiný formát přístupového kódu. Ten začíná dvěma písmeny následovanými podtržítkem a šesti nebo více číslicemi. Například:

- NM_123456 mRNA
- NP_123456 proteiny
- NC_123456 chromozomy

Verze (Version)

```
VERSION     NC_000005.10  GI:568815593
```

Tento kód vychází z kódu přístupového a pouze ho rozšiřuje o informaci o změně v záznamu sekvence. Kód má například tvar U12345.1. Při změně i jediného nukleotidu se zvyšuje číslo za tečkou U12345.2, ale přístupový kód před tečkou zůstává stejný.

GI

Dále se zavádí i další numerický systém GI (GenInfo Identifier). Tento systém přiřazuje jednotlivým záznamům pouze jedno konkrétní číslo. Tento kód nemá žádnou podobnost s kódem přístupovým. A při změně v sekvenci dostává nová verze záznamu nové GI číslo.

GI systém fungoval po mnoho let v NCBI pro sledování historie záznamů v GenBank. V roce 1999 byl zaveden systém přístupového kódu v kombinaci s číslem změny. Dnes běží oba tyto systémy paralelně a při změně záznamu se mění oba kódy.

Klíčová slova (Keywords)

```
KEYWORDS    RefSeq.
```

Obsahují slova nebo fráze, které charakterizují sekvenci. Toto pole bylo využíváno dříve, ale jelikož nemá jasně definovaný slovník, v dnešní době se většinou nevyplňuje. Výjimku lze udělat na přání autora, ale současně nesmí být uvedená informace redundantní k dalším uvedeným v záznamu, anebo se jedná o speciální typ sekvence (např. EST, STS, GSS, HTG atd.). Případně je uvedeno, že sekvence patří do databáze RefSeq.

Zdroj (Source)

```
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
            Euteleostomi; Mammalia; Eutheria; Euarchontoglires;
            Primates; Haplorrhini; Catarrhini; Hominidae; Homo.
```

Toto pole obsahuje zkrácený název organismu, ze kterého záznam pochází. Někdy je tato informace doplněna i o typ molekuly. Nebo jsou zde uvedeny i běžně zavedené názvy organismů např. human, mouse, chimpanzee atd.

Organismus (Organism)
Na prvním řádku v tomto poli je uvedeno přírodopisné jméno v latině (rodové a druhové jméno). Na dalším řádku je potom uvedeno taxonomické zařazení organismu.

Reference

```
REFERENCE   2  (bases 1 to 3824)
  AUTHORS   Schmutz,J., Martin,J., Terry,A., Couronne,O.,
            Grimwood,J. …
  TITLE     The DNA sequence and comparative analysis of human
            chromosome 5
  JOURNAL   Nature 431 (7006), 268-274 (2004)
   PUBMED   15372022
```

Reference obsahují seznam publikací, které pojednávají o datech uvedených v záznamu. Reference jsou automaticky setříděny podle data od nejstarší po nejmladší. U sekvencí, které nebyly prozatím zveřejněny v tištěné podobě, bývá uveden status „unpublished“ nebo „in press“ (nezveřejněno, v tisku). Citace uvedená v referencích může být článek v odborném časopise, kapitola z knihy, kniha, diplomová práce, patent atd. Poslední citace v referencích obvykle slouží jako informace o autorovi daného příspěvku. A proto bývá místo názvu publikace uvedeno „Direct Submission“. U některých starších příspěvků ovšem tato poslední citace chybí.

Autoři (Authors)
Toto pole v citaci obsahuje seznam autorů v pořadí, v jakém jsou uvedeni v citovaném článku.

Název (Title)
Tady je uveden název publikace a u dosud nepublikovaných prací je uveden alespoň předběžný název.

Časopis (Journal)
Zde nalezneme zkrácený název časopisu uvedený v MEDLINE (Medical Literature Analysis and Retrieval System Online).

PubMed
Publikace, které jsou uveřejněny v PubMed, zde mají uvedeno své přístupové číslo do PubMed. A naopak články v PubMed obsahují odkazy na korespondující záznamy sekvencí v GenBank.

Vlastnosti (Features)
Tato část záznamu informuje o genech, jejich produktech a oblastech jejich výskytu v sekvenci. Případně popisuje oblasti sekvence, které kódují proteiny a molekuly RNA.

Zdroj (Source)

```
source          1..3824
                /organism="Homo sapiens"
                /mol_type="genomic DNA"
                /db_xref="taxon:9606"
                /chromosome="5"
```

Toto pole je povinné a shrnuje všechny podstatné informace o záznamu jako je délka sekvence, jméno organismu, ID taxonu, případně číslo chromozomu, na kterém se nalézá daná sekvence nebo typ molekuly. Dále také může obsahovat další podrobnosti zadané autorem příspěvku.

Taxon ID je unikátní identifikační číslo pro taxon zdrojového organismu. Podrobné taxonomické zařazení potom lze dohledat v NCBI Taxonomy Database.

Gen, molekula RNA

```
gene            1..3824
                /gene="EGR1"
mRNA            join(1..577,1266..3824)
                /gene="EGR1"
```

Informuje o přítomnosti genů v dané sekvenci nebo molekuly RNA.

CDS (Coding Sequence)

```
CDS             join(271..577,1266..2590)
                /gene="EGR1"
                /gene_synonym="AT225; G0S30; KROX-24; NGFI-A; TIS8;
                ZIF-268; ZNF225"
                /note="Derived by automated computational
```

```
analysis using gene prediction method: BestRefSeq."
/codon_start=1
/product="early growth response protein 1"
/protein_id="NP_001955.1"
/translation="MAAAKAEMQLMSPLQISDPFGSFPHS...
```

Kódující sekvence (CDS) je úsek nukleotidů, které odpovídají sekvenci aminokyselin v proteinu. Tento úsek je dán číselným intervalem, který má různý formát podle úplnosti CDS.

- Kompletní CDS: *n..m*
- Spojení více úseků CDS: *join (n..m, o..p,...)*
- Neúplné na 5‘ konci : <*n..m*
- Neúplné na 3‘ konci: *n..m*>
- Komplementární vlákno: *complement (n..m)*

Poté mohou být uvedeny další informace například název genu, název organismu, název produktu, jeho funkci a spoustu dalších informací.

Dále v tomto poli nalezneme i identifikační číslo daného proteinu. Toto číslo je podobné číslu přístupovému. Kód se skládá ze tří písmen následovaných pěti číslicemi, tečkou a číslem verze. Při jakékoliv změně sekvence aminokyselin se opět změní i číslo verze při současném zachování kódu před tečkou (AAA12345.1 se změní na AAA12345.2).

Také číselný systém GI (GenInfo Identifier) je zde zaveden, i když v tomto případě je jedinečné číslo přiřazeno translaci proteinů a také se mění stejně jako GI u sekvence nukleotidů.

Poté následuje sekvence aminokyselin translatovaných z odpovídající kódující sekvence nukleotidů.

Původ (Origin)

```
ORIGIN
1   gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg
61  ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct
121 ctgcatctga agccgctgaa gttctactaa gggtggataa catcatccgt gcaagaccaa
```

Posledním údajem v záznamu je samotná sekvence. Sekvence je zapsána tak, aby na jednom řádku bylo šedesát nukleotidů. Ty jsou dále děleny do šesti sloupečků, které jsou od sebe odděleny mezerou a vytváří tak skupiny deseti nukleotidů. Každý řádek začíná číslem, které vyjadřuje pořadí prvního nukleotidu na řádku v celé sekvenci.

1.2.2. FASTA

Datový formát FASTA redukuje množství informací pouze na základní údaje. Skládá se ze dvou částí: hlavičky a sekvence [3].

```
   >gi|171361|gb|L04459|YSCCYS3A    Saccharomyces    cerevisiae
cystathionine gamma-lyase (CYS3) gene, complete cds.

GCAGCGCACGACAGCTGTGCTATCCCGGCGAGCCCGTGGCAGAGGACCTCGCTTGCGAAAGCATCGAGTACC
GCTACAGAGCCAACCCGGTGGACAAACTCGAAGTCATTGTGGACCGAATGAGGCTCAATAACGAGATTAGCG
ACCTCGAAGGCCTGCGCAAATATTTCCACTCCTTCCCGGGTGCTCCTGAGTTGAACCCGCTTAGAGACTCCG
AAATCAACGACGACTTCCACCAGTGGGCCCAGTGTGACCGCCACACTGGACCCCATACCACTTCTTTTTGTT
ATTCTTAAATATGTTGTAACGCTATGTAATTCCACCCTTCATTACTAATAATTAGCCATTCACGTGATCTCA
GCCAGTTGTGGCGCCACACTTTTTTTTCCATAAAAATCCTCGAGGAAAAGAAAAGAAAAAAATATTTCAGTT
ATTTAAAGCATAAGATGCCAGGTAGATGGAACTTGTGCCGTGCCAGATTGAATTTTGAAAGTACAATTGAGG
CCTATACACATAGACATTTGCACCTTATACATATAC
```

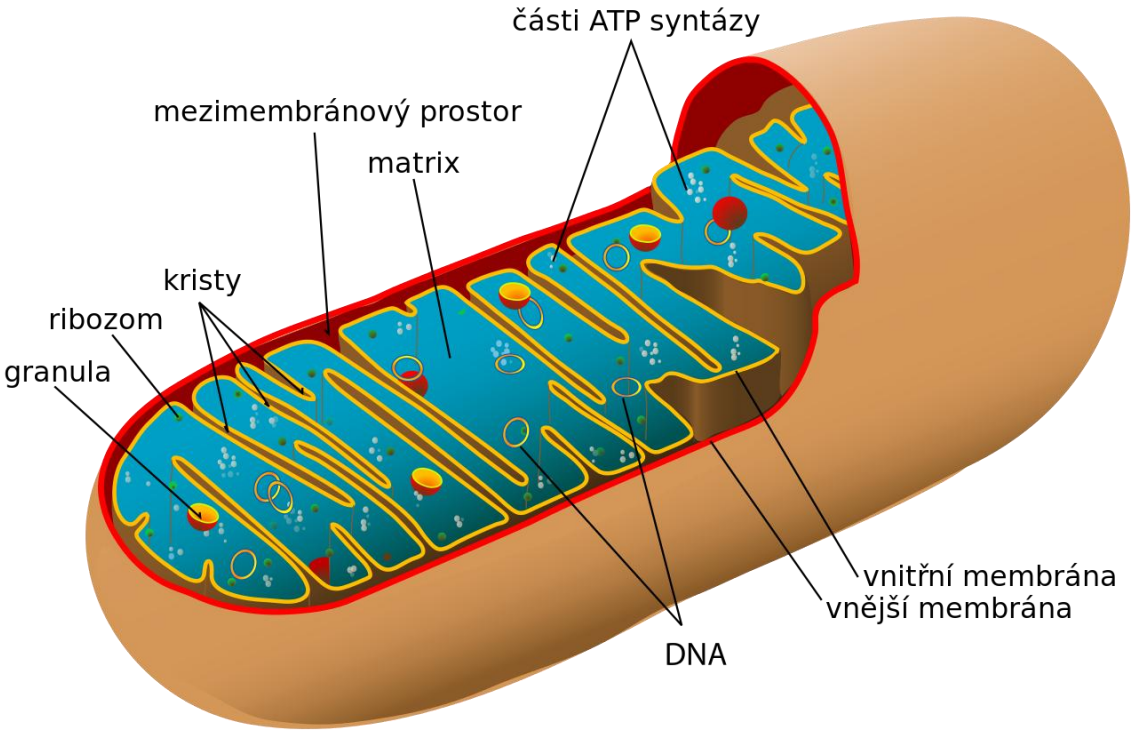
Hlavička každého FASTA soubor má na prvním řádku znak „>“, ten značí začátek nového souboru. U nejjednodušších souborů FASTA následuje identifikační kód (*Locus Name*) a poté na druhém řádku již samotná sekvence (obvykle se šedesáti znaky na řádek). U modifikovaných souborů (viz výše) bývá v hlavičce uvedeno *GI* číslo, *Locus Name* a pole *Definice* z GenBank záznamu. Dále je opět uvedena příslušná sekvence.

2. Mitochondriální a plastidová DNA

2.1. Mitochondrie

2.1.1. Stavba

Mitochondrie jsou organely eukaryotických buněk. Uvnitř buňky můžeme mitochondrií nalézt několik set až sto tisíc v závislosti na typu buňky. Mitochondrie mají podlouhlý případně kulovitý tvar a jsou veliké jeden až několik mikrometrů. Jsou složené ze dvou membrán.



Obrázek 1 Stavba mitochondrie
(Převzato z: http://cs.wikipedia.org/wiki/Mitochondrie)

Vnější membrána je hladká a má podobné vlastnosti jako buněčná membrána eukaryotické buňky. Je propustná zejména pro ionty a je tvořena převážně lipidy.

Vnitřní membrána je zvrásněná a uvnitř mitochondrie vytváří kristy nebo trubičky. Což jsou záhyby vnitřní membrány, které probíhají napříč mitochondrií. Díky tomuto zvrásnění je obsah plochy vnitřní membrány několika násobně větší. Mitochondrie s kristami nalezneme ve většině buněk v lidském těle. Mitochondrie s vnitřní membránou v podobě trubiček lze najít u buněk syntetizujících steroidy (buňky žlutého tělíska, kůry nadledvin, varlete). Membrána je z malé části tvořena lipidy, větší část je tvořena proteiny dýchacího řetězce, a tudíž je membrána málo propustná pro ionty. Uvnitř vnitřní membrány mitochondrie nalezneme mitochondriální matrix. V ní jsou

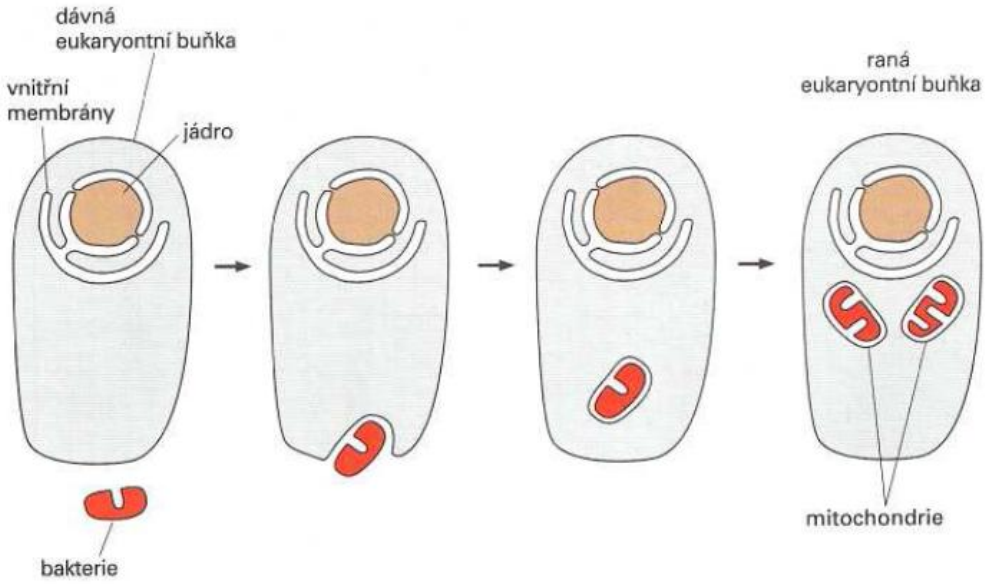
obsaženy molekuly mitochondriální DNA a RNA, enzymy dýchacího řetězce, intramitochondriální granula, ribozomy… [8]

2.1.2. Funkce

Mitochondrie produkují hlavní zdroje energie pro buňku a to molekuly ATP (adenosintrifosfátu). Energie se získává pomocí několika biologických oxidací. Během těchto oxidací se v prvním kroku z organických látek uvolní aktivní vodík (citrátový cyklus), který je dále v dýchacím řetězci postupně oxidován až na vodu. Při této oxidaci dochází k přenosu elektronů přes mitochondriální membránu pomocí přenašečů a současně probíhá oxidativní fosforylace, která fixuje uvolněnou energii v podobě makroenergetických vazeb v molekulách ATP. Celý proces se nazývá buněčné dýchání, protože je během těchto oxidací spotřebováván kyslík a uvolňuje se oxid uhličitý[1].

2.1.3. Mitochondriální DNA

Původ mitochondrií není jednoznačně potvrzen, ovšem existují mnohé teorie na toto téma. Předpokládá se, že mitochondrie byla v minulosti samostatně žijící prokaryotický organismus, který v průběhu evoluce splynul s vyvíjející se eukaryotickou buňkou a vytvořil s ní tak endosymbiotický vztah. Vztah aerobního organismu s anaerobním. Podobně u rostlin pravděpodobně vznikly i chloroplasty a další plastidy. Teorie potvrzuje i fakt, že u všech zmíněných organel byla nalezena vlastní mimojaderná DNA a navíc stavbou spíše připomínají prokaryotické organismy.



Obrázek 2 Endosymbiotická teorie vzniku mitochondrie
(Převzato z [8])

U mitochondrií se také předpokládá, že přítomnost dvojité membrány je též způsobena endosymbiózou. Vnější membrána patřila kdysi anaerobnímu organismu a

vnitřní membrána je pozůstatkem kdysi samostatného aerobního organismu. V dnešní době je ovšem existence samostatné mitochondrie vyloučena. Během vývoje eukaryotické buňky došlo k výraznému zkrácení mitochondriální DNA (mtDNA) a část genů byla přesunuta do jaderné DNA.

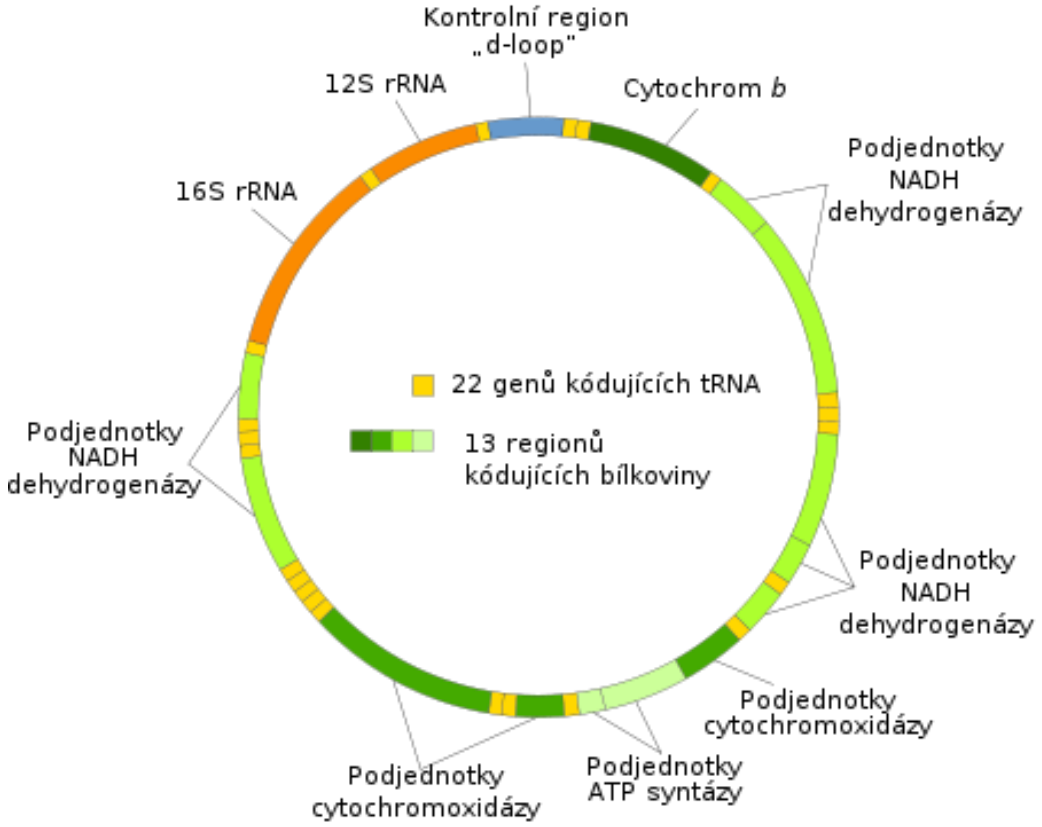
Díky svému původu má i mtDNA prokaryontní charakter, cirkulární tvar (na rozdíl od eukaryotické DNA, která má tvar lineární). Další odlišností od jaderné DNA je i fakt, že se v mitochondrii mtDNA objevuje v několika kopiích, přičemž buňka samotná obsahuje velké množství mitochondrií. Dále má mitochondrie i vlastní ribozomy pro syntézu proteinů z mRNA. Další rozdíl je i v kódování aminokyselin u mtDNA. Například stop-kodon u jaderné mRNA UGA v mitochondriální DNA odpovídá kodonu pro aminokyselinu tryptofan. Naopak u mtRNA nalézáme nové stop-kodony AGA, AGG, které u jaderné DNA vytváří kodon pro animokyselinu arginin [9].

Mitochondriální DNA živočichů

Mitochondriální DNA a je nositelem mimojaderné dědičnosti. Tato DNA je u živočichů dědičná pouze po matce, jelikož mitochondrie otce jsou při oplození zničeny a ve vajíčku zůstanou pouze mitochondrie matky [13].

Délka mitochondriální DNA se liší podle živočišného druhu a během evoluce byla výrazně zkrácena. Její délka bývá průměrně 16-19kb a u živočichů mtDNA neobsahuje žádné introny. U člověka má mtDNA délku 16 569 bp, které kódují 37 genů [16].

- 2 geny kódují rRNA: s-rRNA či 12S rRNA (small rRNA, „malá“ podjednotka rRNA) a l-rRNA či 16S r-RNA (large rRNA, „velká“ podjednotka rRNA)
- 22 genů kóduje 22 druhů molekul tRNA
- 13 genů kóduje proteiny pro oxidační fosforylaci:
    - ND1 (gen NADH dehydrogenázy podjednotky 1)
    - ND2 (gen NADH dehydrogenázy podjednotky 2)
    - COX1 (gen cytochromoxidázy podjednotky 1)
    - COX2 (gen cytochromoxidázy podjednotky 2)
    - ATP8 (gen ATP syntázy podjednotky 8)
    - ATP6 (gen ATP syntázy podjednotky 6)
    - COX3 (gen cytochromoxidázy podjednotky 3)
    - ND3 (gen NADH dehydrogenázy podjednotky 3)
    - ND4L (gen NADH dehydrogenázy podjednotky 4L)
    - ND4 (gen NADH dehydrogenázy podjednotky 4)
    - ND5 (gen NADH dehydrogenázy podjednotky 5)
    - ND6 (gen NADH dehydrogenázy podjednotky 6)
    - CYTB (gen cytochrom b)



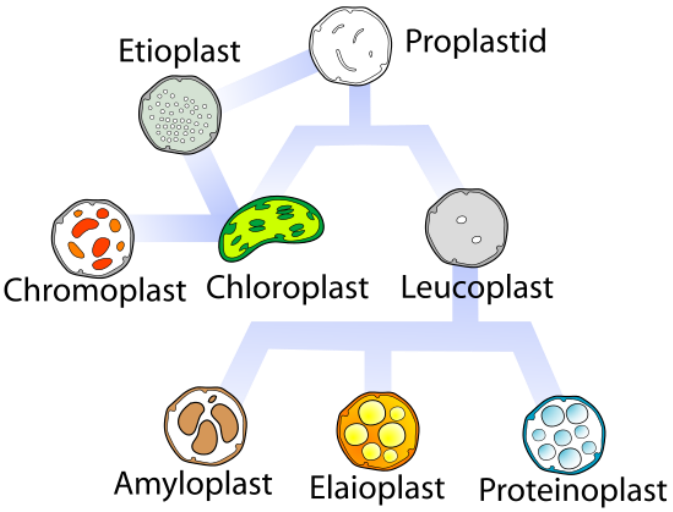
Obrázek 3 Mitochondriální DNA člověka
(Převzato z: http://cs.wikipedia.org/wiki/Mitochondrialni_DNA)

Mitochondriální DNA rostlin

U vyšších rostlin může být mtDNA delší než u živočichů. U rostlin také nalézáme kromě kruhovité mtDNA také lineární úseky. Stejně jak u živočichů i v rostlinných mitochondriích nalézáme mnoho kopií mitochondriálního genomu. Ovšem tyto fakta se různí druh od druhu (délka sekvence může být 200 nebo i 2500 kpb). I přesto, že některé mitochondriální genomy jsou o mnoho větší než u živočichů, nekódují všechny potřebné geny a molekuly RNA. Část těchto informací byla opět přenesena do jaderné DNA. Nicméně u rostlin bylo identifikováno 57 známých proteinů a mnohé další předpokládané ORF [14].

2.2. Plastidy

Plastidy jsou významné eukaryotické organely rostlin, které stejně jako mitochondrie pravděpodobně vznikly díky endosymbióze. Během evoluce se plastidy staly základními stavebními kameny pro život a funkci rostliny. Zejména díky jejich schopnosti vytvořit konkrétní funkční jednotku podle potřeby rostliny [10].



Obrázek 4 Typy plastidů
(Převzato z: http://en.wikipedia.org/wiki/Plastid)

- **Proplast**

  Všechny plastidy vznikají z malých nediferenciovaných proplastidů, které lze nalézt ve vyvíjecích se buňkách meristému (dělivého pletiva). Během diferenciace se proplastid vyvine podle typu sousedícího plastidu.

- **Etioplast**

  Tyto plastidy jsou druhem chloroplastů, které dosud nebyly vystaveny slunci (například pupeny nebo základy listů). V etioplastech již nalezneme základy tylakoidů a prolamelární těleso s protochlorofylem, ze které může vzniknout chlorofyl.

- **Chloroplast**

  Chloroplast je specializovaný plastid schopný fotosyntézy. Má rozmanitou velikost i tvar. Od cytoplazmy je oddělen dvěma membránami, které se liší složením (obsahem lipidů i proteinů). Hustý roztok plný enzymů uvnitř chloroplastu se nazývá stroma. V něm se nachází další membránové struktury – tylakoidy. Tyto ploché útvary se mohou seskupovat a vytvářet tzv. grana. Dále zde můžeme nalézt i tylakoidy integranální nebo stromatální, které jednotlivá grana spojují. Uvnitř tylakoidu se nachází lumen.

  Absorpce slunečního záření probíhá v membráně tylakoidů. Tato energie je využita na štěpení vody za uvolnění molekuly kyslíku a protony (H^+) se hromadí v tylakoidu. Tento protonový gradient poté slouží jako energetický zdroj při syntéze ATP. Elektrony uvolněné při štěpení vody jsou využity na vnější straně tylakoidu k tvorbě NADPH. V další fázi („temnostní fázi“) jsou tyto produkty využity k tvorbě sacharidů z oxidu uhličitého [8].

- **Chromoplast**

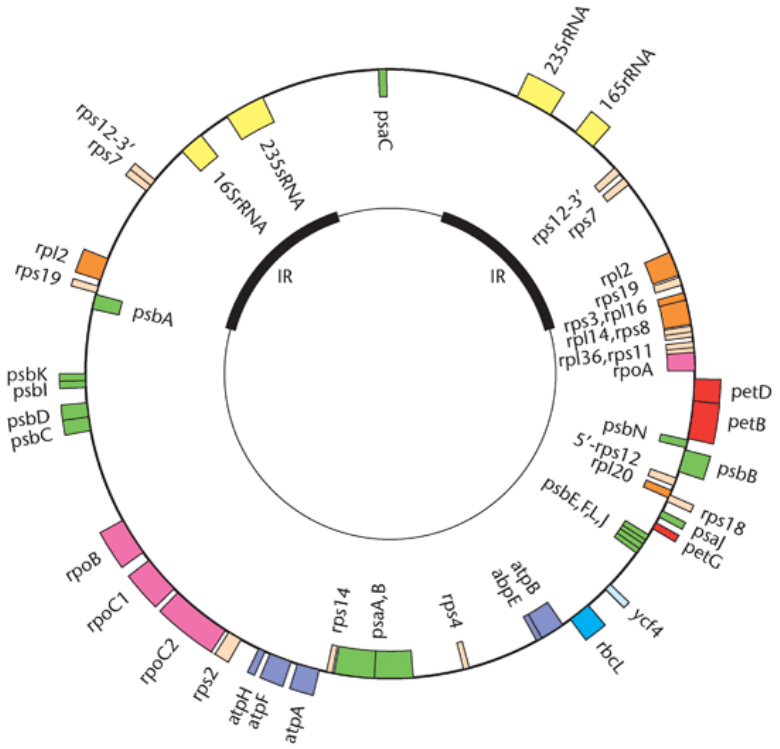
  Chromoplast je žlutý nebo červený plastid, který již není schopný fotosyntézy. Vznikl ze starého chloroplastu ztrátou chlorofylu. Chromoplast syntetizuje pigment (např. karotenoid) a může mít i zásobní funkci. Lze ho nalézt na podzim v listech, kde způsobuje typické vybarvení listů v tomto ročním období (listy přichází o chlorofyl a odhalí se tak další pigment v plastidu). Anebo je součástí plodů rostlin a dodává jim výrazné barvy. Tím dokáže přilákat pozornost konzumentů, kteří rozšiřují semena plodů dál. Chromoplasty najdeme i u kořenové zeleniny jako je mrkev nebo sladký brambor, kde slouží jako zásobárny látek nerozpustných ve vodě.

- **Leukoplast**

  Leukoplasty jsou bezbarvé plastidy nacházející se ve především v kořenech rostlin a v částech, které se nepodílejí na fotosyntéze. Jejich význam je především v zásobní funkci. Proteinoplasty shromažďují proteiny, amyloplasty škrob a elaioplasty oleje.

2.2.1. Plastidová DNA

Rostlinné buňky kromě jaderného a mitochondriálního genomu mají též genom chloroplastový (plastidový) – cpDNA. Chloroplastová DNA se stejně jako mtDNA vyskytuje v mnoha kopiích, počet závisí na typu buňky a druhu organismu.



Obrázek 5 Plastidová DNA
(Převzato z: [15])

Vyšší rostliny mívají 20 až 200 chloroplastů, ve kterých najdeme 10 až 50 kopií cpDNA. Co se týče velikosti genomu, pohybuje se mezi 35 a 220 kpb a u fotosyntetizujích rostlin mezi 115 a 217 kpb. Na cpDNA nalezneme geny pro proteiny a geny pro strukturální RNA. Celkem je to 54 až 259 genů a u vyšších rostlin je to kolem 120 genů [14]. Dříve se předpokládalo, že cpDNA je pouze kruhovitého tvaru. Tvoří ji pár invertovaných repetic (kopie sekvence DNA mající opačnou orientaci) různé délky. Dnes již víme, že převažují zřejmě lineární větvené komplexní molekuly DNA [2].

Chloroplastová DNA dědičnost je většinou uniparentální, matroklinní a byla zaznamenána u jehličnanů a některých krytosemenných rostlin. Nicméně se může vyskytnout i dědičnost biparentální.

3. Extrakce dat z GenBank formátu

3.1. Bioinformatický toolbox

Programové prostředí Matlab lze využít analýze biologických dat. K tomu slouží bioinformatický toolbox, který lze do Matlabu doinstalovat. Tento toolbox obsahuje algoritmy a aplikace pro analýzu microarray (pro načtení, filtrování, normování a vykreslení), hmotnostní spektrometrii (předzpracování, klasifikace a identifikace markerů), ontologie genů a další. Funkce v toolboxu umožňují načítání dat z různých datových formátů např. SAM, FASTA,CEL, GenBank. A následně dokáže sekvence prohledávat, vizualizovat, více násobně zarovnávat, vytvářet fylogenetické stromy [11].

3.1.1. Funkce genbankread

Pro načtení souborů v datovém formátu GenBank slouží v Matlabu funkce `genbankread`. Zapsaná například takto:

```
GenBankData = genbankread(File)
```

Argument *File* (soubor) může být název souboru, cesta k souboru a jeho název, nebo URL odkazující na daný soubor. V případě, že je zadán pouze název souboru, musí se tento soubor nalézat v právě používané složce. Dále může být argument *File* i text odpovídající zápisem struktuře datového formátu GenBank.

Argument *GenBankData* je struktura, která vznikne načtením souboru. Jednotlivé pole *fields* této struktury odpovídají klíčovým slovům GenBank a v poli *values* najdeme získané informace. Ty jsou uvedeny ve formě znakového řetězce anebo dalších „podstruktur“ (v případě, že se klíčové slovo objevuje vícekrát v záznamu, vytvoří se tyto pod „podstruktury“ například u položky *Reference* a *CDS*).

```
GenBankData =

              LocusName: 'NC_004537'
     LocusSequenceLength: '15076'
    LocusNumberofStrands: ''
          LocusTopology: 'circular'
       LocusMoleculeType: 'DNA'
    LocusGenBankDivision: 'INV'
   LocusModificationDate: '01-FEB-2010'
              Definition: 'Branchiostoma belcheri mitochondrion, complete genome.'
              Accession: 'NC_004537'
                Version: 'NC_004537.1'
                     GI: '27802003'
                Project: []
                 DBLink: 'Project: 12118 BioProject: PRJNA12118'
               Keywords: 'RefSeq.'
                Segment: []
                 Source: 'mitochondrion Branchiostoma belcheri (amphioxus)'
         SourceOrganism: [3x64 char]
              Reference: {[1x1 struct]  [1x1 struct]  [1x1 struct]}
                Comment: [3x64 char]
               Features: [284x74 char]
                    CDS: [1x13 struct]
               Sequence: [1x15076 char]
```

Obrázek 6 Příklad načteného souboru pomocí funkce `genbankread`

Spolehlivost funkce `genbankread` byla testována na datech získaných z online databáze GenBank. Tyto data byly kompletní mitochondriální sekvence a pro vyhodnocení úspěšnosti funkce bylo použito 3 873 souborů. Tyto soubory byly načteny do Matlabu pomocí funkce `genbankread` a uloženy do jediné struktury, která byla dále analyzována.

Při samotném načítání nedošlo k žádným chybám, na které by upozornila hláška v *Command Window* a výsledná struktura obsahovala všech 3 873 načtených souborů. Za předpokladu, že funkce vykazuje chybovost, byla zkontrolována povinná pole souboru. Zvláště byla pozornost soustředěna na pole *CDS* a *Sequence*, které pro nás mají největší význam.

Pomocí nově vytvořené funkce byla zkontrolována přítomnost sekvence, což se potvrdilo a následně byla i ověřena kompletnost této sekvence. Porovnáním hodnoty v poli *LocusSequenceLength* s délkou znakového řetězce v poli *Sequence* a bylo potvrzeno úspěšné načtení sekvencí u všech souborů.

Dále se ověřilo pole *CDS*. U třech souborů ovšem toto pole nebylo nalezeno. Pro ověření správnosti načtení byly k těmto záznamům dohledány i původní příspěvky v internetové databázi. V případě prvního příspěvku byl nalezen tento záznam *Features*:

```
FEATURES             Location/Qualifiers
     source          1..3084
                     /organism="Alatina moseri"
                     /organelle="mitochondrion"
                     /mol_type="genomic DNA"
                     /isolate="A"
                     /db_xref="taxon:675638"
                     /chromosome="5"
                     /country="USA: Waikiki, Honolulu, Oahu, Hawaii"
     repeat_region   complement(1..703)
                     /function="telomere"
                     /rpt_type=inverted
                     /rpt_type=terminal
     gene            711..3084
                     /gene="rnl"
                     /db_xref="GeneID:11946160"
     rRNA            711..2389
                     /gene="rnl"
                     /product="large subunit ribosomal RNA"
                     /db_xref="GeneID:11946160"
     repeat_region   2382..3084
                     /function="telomere"
                     /rpt_type=inverted
                     /rpt_type=terminal
```

Obrázek 7 Pole Features u Alatina moseri mitochondrion chromosome 5, complete sequence

(Získáno z: http://www.ncbi.nlm.nih.gov/nuccore/NC_017011)

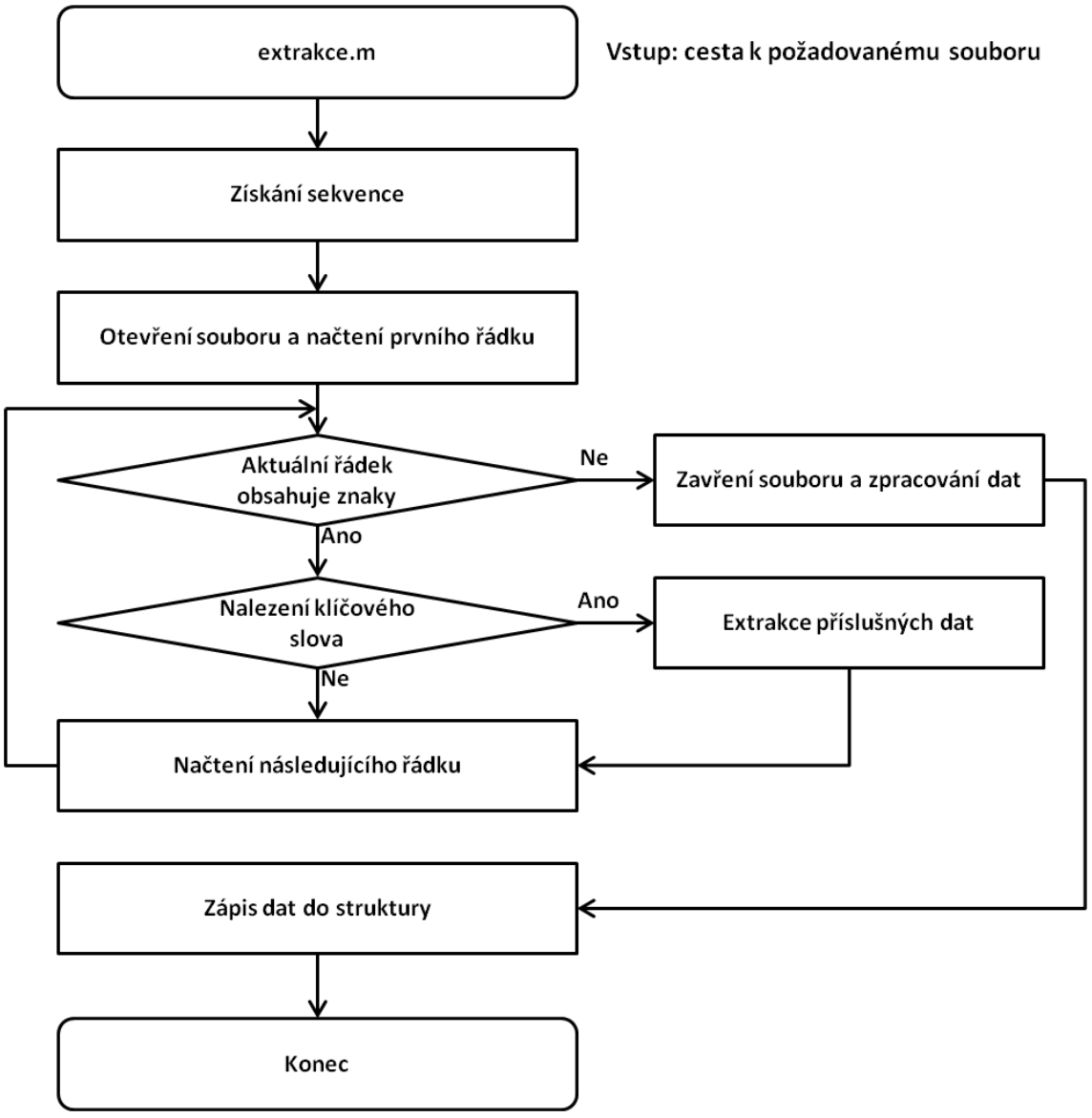
Z tohoto záznamu je patrné, že autor samotného příspěvku pole *CDS* vůbec neuvedl. Zde vyplívá na povrch problém korektnosti všech záznamů v databázi, kdy autor příspěvku nemusí dodržet předem danou normu záznamu. A v takovém případě mohou být podstatné informace funkcí `genbankread` ignorovány.

4. Programové řešení

V rámci programového řešení byl vyvinut program na automatickou extrakci dat z GenBank souborů, který je alternativou k funkci `genbankread` z bioinformatického toolboxu. Pro vhodnou reprezentaci dat byla vytvořena uživatelská aplikace, která získaná data vypíše a umožní uložení těchto dat do FASTA souboru.

4.1. Funkce pro extrakci dat z GenBank souboru

Základem uživatelské aplikace je funkce `extrakce.m`, která pracuje podle níže uvedeného vývojového diagramu a byla vytvořena na základě informací z kapitoly 1.2.1.



Obrázek 8 Vývojový diagram funkce extrakce.m

Vstupem této funkce je cesta k souboru, který chceme extrahovat a jejím výstupem je struktura extrahovaných dat. Samotná funkce se skládá z následujících funkčních bloků:

Získání sekvence
Pro získání sekvence z GenBank souboru byla napsána nová funkce `sekvence.m`. Vstupem této funkce je cesta k souboru a výstupem je extrahovaná sekvence, její délka, výpis jiných znaků než A, C, G, T a jejich pozice v sekvenci.

Tato funkce v prvním kroku načte všechny řádky do jediné proměnné pomocí funkce `fgetl`. Po té hledá klíčové slovo *ORIGIN*, které značí počátek zápisu sekvence v souboru, a extrahuje veškerý text za tímto slovem. Tento text ovšem obsahuje také mezery, číselná označení pozice bází a znak označující konec sekvence. V dalším kroku se tyto znaky eliminují a výsledkem je kompletní sekvence.

I tato sekvence ovšem může obsahovat další znaky kromě předpokládaných znaků (A, C, G, T), které mohly vzniknout v důsledku špatné sekvenace. Tyto znaky jsou detekovány a uloženy do nové proměnné (včetně pozice a příslušného znaku), která je následně použita v uživatelské aplikaci.

Otevření souboru
Soubor je otevřen pomocí funkce `fopen`, uloží se identifikátor souboru a načte se první řádek souboru.

Cyklus pro čtení jednotlivých řádků
Pro získání potřebných dat je nutné celý soubor procházet řádek po řádku a postupně jej analyzovat. Požadovaná data jsou v souboru uložena pod klíčovými slovy. V tomto případě jsou podstatná pole: *DEFINITION*, *ACCESSION*, *ORGANISM*, *D-loop*, *CDS*, *rRNA*, *tRNA*.

Nalezení klíčových slov
K nalezení klíčových slov byla použita funkce `strfind`, která na svém výstupu uvede pozici znaku, kde hledané slovo nebo znakový řetězec začíná.

Při nalezení slova *DEFINITION* očekáváme, že nalezneme podrobnosti o čteném souboru, zejména o typu genomu. Jelikož je software určen pro čtení mitochondriálních a plazmidových genomů. Nalezený typ genomu je uložen do proměnné.

Slovo *ACCESSION* označuje řádek, na kterém nalezneme přístupový kód souboru v GenBank databázi. A slouží k jednoznačnému označení souboru při extrakci dat do FASTA souboru – je uvedeno v hlavičce souboru, spolu s názvem organismu, který se nachází na stejném řádku jako klíčové slovo *ORGANISM*.

Dále jsou vyhledávány kódující úseky na sekvenci (D-loop, CDS, rRNA, tRNA). Při jejich nalezení je nutné ověřit i pozici, na které bylo dané slovo nalezeno, jelikož se může vyskytnout i v anotaci souboru. V položce *Features* tato slova vždy začínají na šestém znaku. Při splnění této podmínky již dochází k extrakci dat a to pomocí funkce `extrakce_sek.m`, která byla za tímto účelem vytvořena.

Tato funkce slouží k extrakci číselných pozic, na kterých se nachází příslušný kódující úsek, a získání odpovídající části sekvence. Na vstupu je nutné uvést aktuální řádek, identifikátor souboru a sekvenci. Při získávání těchto číselných pozic bylo nutné zohlednit několik možností jejich zápisu.

```
CDS             join(74631..74744,complement(103553..103784),
                complement(102987..103012))
```

Zaprvé bylo potřeba zjistit, zda toto pozice není uvedena na více řádcích souboru. Proto se v prvním kroku načte další řádek souboru a ověří se, zda zde zápis pokračuje či nikoliv. Pokud ano zapíše se s původním řádkem do další proměnné, načte se následující řádek a znova se ověří. Po úspěšném zapsání všech řádků do jediné proměnné je zjištěno, zda obsahuje tato proměnná slova *join* a *complement*. *Join* značí, že kódující úsek je rozdělen na více částí a *complement* říká, že kódující úsek se nachází na komplementárním vlákně.

Pro úspěšnou extrakci sekvence je nutné zjistit v případě výskytu obou výrazů, které slovo je uvedeno jako první. V případě, že prvním slovem je *complement*, extrahuje se označení pozice a odpovídající úsek sekvence. Jednotlivé úseky sekvence se řadí za sebe a po zapsání všech částí se získá příslušný reverzní komplement celého úseku. Reverzní komplement je vytvořen pomocí funkce `ReverseKomplement.m`, která vytvoří komplement ke vstupní sekvenci a následně provede reverse tohoto komplementu.

Pokud je první slovo *join* může zápis vypadat podobně, jako bylo uvedeno výše. Což znamená, že pouze některé úseky se nacházejí na komplementárním vlákně. V této situaci je nutné nejprve určit, kolik takových úseků v proměnné existuje a kde se nachází. Následně tyto pozic extrahovat a ukládat včetně příslušného komplementárního úseku sekvence. Na závěr se extrahují zbylé úseky, které se na sekvenci nachází v obvyklém směru na hlavním vlákně. V posledním kroku se všechny získané úseky sekvence poskládají za sebe v pořadí, které je uvedeno ve výchozí proměnné.

Výsledkem této funkce tedy bude/budou dvojice čísel označující začátek a konec kódujícího úseku, sekvence nacházející se na tomto úseku a poslední načtený řádek. Tato získaná data jsou uložena do proměnných v závislosti na nalezeném klíčovém slově. V případě, že se jednalo o CDS lze extrahovat i název toho genu, jelikož se

nalézá na prvním řádku pod zápisem pozic. A v případě, že se jednalo o některé z RNA, extrahuje se kromě názvu i produkt.

Zavření souboru a zpracování dat

Po zkontrolování všech řádků souboru se aktuální soubor uzavře a získaná data se dále zpracují.

Nejdříve data získaných kódujících úseků (D-loop, CDS, tRNA, rRNA) vstupují do funkce `serazeni.m`. Tato funkce má za úkol nejprve zapsat všechna data do jediné proměnné a následně tato data seřadit pomocí funkce `sortrows` vzestupně podle prvního sloupce, kde je uvedena počáteční pozice daného úseku.

Následně se tyto získaná data použijí ve funkci `nepouzite.m` ke zjištění, které části sekvence nepatří ani do jednoho z nalezených úseků. Tato informace je velmi důležitá vzhledem k faktu, že na mitochondriální DNA by se neměly vyskytovat nekódující úseky (viz 2.1.3). Ke zjištění těchto úseků se vypočítává rozdíl mezi pozicí označující konec kódujícího úseku a pozicí označující začátek následujícího úseku. Získaná data se uloží do nové proměnné.

Pro lepší zobrazení dat v uživatelském rozhraní ještě obě výše zmíněné funkce provedou zápis pozic do jedné buňky v prvním sloupci. U `nepouzite.m` do druhého sloupce uvede délku toho úseku. U `serazeni.m` se do druhého sloupce pouze přepíše název kódujícího úseku.

Zápis dat do výstupní struktury

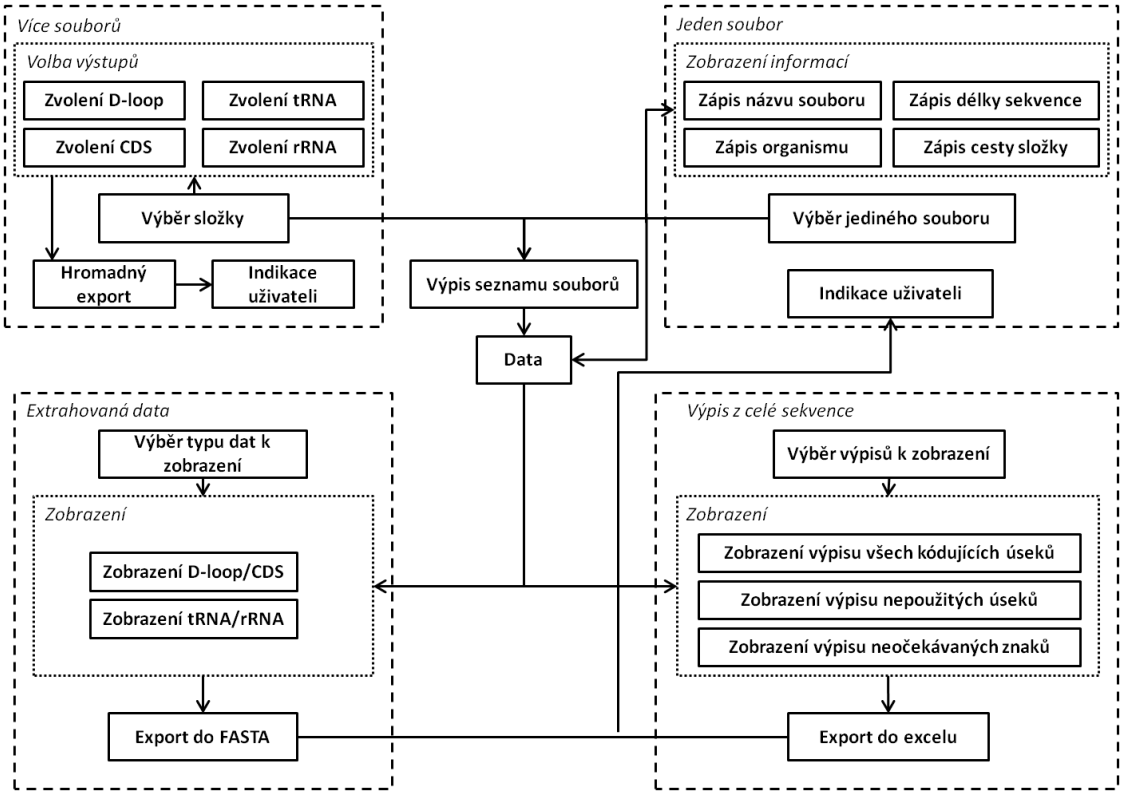
Výstupem funkce `extrakce.m` je struktura *Data*, do které se zapíší všechny výše zmíněná získaná data.

```
Data =

            Nazev: 'NC_012920'
       Organismus: 'Homo sapiens'
              Typ: 'Mitochondrion'
    DelkaSekvence: 16569
         Sekvence: [1x16569 char]
            Pozor: []
            Dloop: {2x4 cell}
              CDS: {13x4 cell}
             tRNA: {22x5 cell}
             rRNA: {2x5 cell}
      SrovnaneGUI: {39x2 cell}
         Srovnane: {39x3 cell}
     NepouziteGUI: {11x2 cell}
        Nepouzite: {11x3 cell}
```

4.2. Uživatelská aplikace

Pro snadnou manipulaci a reprezentaci dat získaných z popsané funkce `extrakce.m` bylo vytvořeno uživatelské rozhraní. Toto rozhraní umožňuje získaná data vhodně zobrazit a případně je exportovat. Celé blokové schéma propojení jednotlivých funkčních bloků je uvedeno na Obrázek 9. Vzhled celé uživatelské aplikace lze nalézt v příloze.



Obrázek 9 Blokové schéma zapojení uživatelského rozhraní

Inicializace

Prvním krokem k použití této aplikace je vybrání GenBank souboru nebo složky a to pomocí tlačítka *„Vyberte soubor“* nebo *„Vyberte složku“*. Obě varianty otevřou uživateli nové okno, kde může vyhledat požadovaný soubor nebo složku.

V případě vybírání pouze jednoho souboru nové okno umožňuje výběr GenBank souborů, soubory s jinou koncovkou než „*.gb“ nebo „*.gbk“ se nezobrazí. Po úspěšném vybrání souboru se provede extrakce dat pomocí funkce `extrakce.m` a získaná data se uloží do proměnné Data. Do panelu *Jeden soubor – zobrazení dat* jsou uvedeny základní informace o souboru, které byly získány z proměnné Data, a to název souboru (odpovídající položce *ACCESION* v originálním souboru), název organismu (položka *ORGANISM*), délka sekvence a cesta k souboru. Dále se vybraný soubor také objeví v listboxu *seznam_souboru* a to i s ostatními GenBank soubory nalézajícími se ve

stejné složce. Pro lepší orientaci se barevně zvýrazní v tomto seznamu původní vybraný soubor.

Při výběru složky se v novém okně pro výběr zobrazí pouze názvy složek, žádné soubory se zde nevypisují. Při úspěšném výběru dojde ke zjištění názvů všech souborů v této složce. Z nich se vyselektují pouze GenBank soubory a jejich názvy se vypíší do listboxu *seznam_souboru*. Pokud se ve vybrané složce žádné GenBank soubory nenachází, uživatel je na tu to skutečnost upozorněn hláškou v listboxu *seznam_souboru*. Pro následné načtení dat pomocí funkce `extrakce.m` a zobrazení těchto dat v pravé části programu je nutné na požadovaný soubor kliknout. Na konec se ještě vypíše cesta ke složce do listboxu *cesta*.

Zobrazení dat

Další možností této aplikace je zobrazení extrahovaných kódujících úseků do tabulky v panelu *„Extrahovaná data“*. K vybrání požadovaných dat je v aplikaci umístěno pop-up menu. V něm má uživatel na výběr položky:

- *D-Loop*
- *CDS*
- *tRNA*
- *rRNA*

Při vybrání jedné z nich dojde k načtení příslušných dat do tabulky. V případě, že se jedná o D-loop a CDS má tabulka čtyři sloupce (*Název*, *Umístění (od), Umístění (do), Sekvence*), u obou RNA tabulka navíc zobrazuje i informaci o produktu. Poslední sloupec, který zobrazuje sekvenci, je nastaven tak, aby při kliknutí na požadovanou sekvenci, došlo k jejímu vypsání do listboxu *sekvence* a uživatel měl tak možnost si celou sekvenci prohlédnout. V případě, že kódující úsek je rozdělen na více částí, do tabulky se vypíší všechny pozice včetně názvu úseku. Nicméně sekvence bude zobrazena jen jedna (vždy u první pozice) a bude se jednat o spojení všech úseků dohromady.

Pro zobrazení zbylých informací o souboru slouží tabulka v panelu *„Výpis z celé sekvence“*. Zde je opět na výběr z pop-up menu:

- **Výpis z celé sekvence** – výpis všech kódujících úseků seřazených podle umístění na sekvenci
  V prvním sloupci zobrazí umístění a ve druhém název položky.
- **Výpis nepoužitých bází**
  V prvním sloupci zobrazí umístění a ve druhém délku tohoto úseku
- **Výpis neočekávaných znaků v sekvenci**

V prvním sloupci zobrazí nalezený znak a ve druhém sloupci umístění na sekvenci. V případě, že se jedná o celý úsek stejného znaku, je tento zápis zjednodušen do jediného řádku a v prvním sloupci je zapsán počátek i konec tohoto úseku.

Export dat

Export dat z jednoho souboru do FASTA souboru

Nedílnou součástí této uživatelské aplikace je i možnost exportu dat. Exportovat lze data v rámci jednoho souboru nebo data ze všech souborů ve zvolené složce.

Pro jeden soubor lze v panelu *„Extrahovaná data“* stisknout tlačítko *„Zapiš ___ do fasta souboru“*. Místo znaků „___“ se zobrazí název kódujících úseků, které jsou aktuálně vybrány v pop-up menu tohoto panelu. Jeho stisknutí inicializuje zápis zobrazených dat v tabulce do souboru ve formátu FASTA (viz kapitola 1.2.2).

Pro zápis do FASTA souboru je potřebná struktura obsahující položky: *Header* (hlavička) a *Sequence* (sekvence). Pro vytváření těchto struktur automaticky byly vytvořeny funkce `zapis_cds.m` a `zapis_RNA.m`. Tyto funkce vezmou všechny data v tabulce (u CDS všech 13 genů, u RNA 22 tRNA nebo 2 rRNA), seřadí je podle abecedy a zapíší do struktury. Hlavička každé sekvence je vytvořena z dat získaných z dřívější extrakce. Zápis může například vypadat takto:

```
>NC_004743 | Acipenser transmontanus | ATP6 | (8154:8836)
```

Jako první je uveden identifikační kód z GenBank databáze, následuje latinský název organismu, název kódujícího úseku a jeho pozice na sekvenci.

Pro zapsání pozice ve výše uvedeném tvaru byla vytvořena funkce `ziskani_pozice.m`. Úkolem této funkce je nalézt všechny pozice týkající se daného kódujícího úseku a zapsat je do jediné buňky tak, aby se daly použít v hlavičce FASTA souboru. Funkce vychází z předpokladu, že v případě, kdy jsou pozice zapsány na více řádcích, tak sekvence je zapsána pouze na prvním řádku. Tyto pozice se poté automaticky sloučí do jediné buňky.

Při zápisu CDS u mitochondriálního genomu se navíc zkontroluje nalezení všech genů. V případě, že některý chybí, zapíše se odpovídající hlavička a do sekvence se zapíše pouze písmeno „n“. Tato skutečnost může být následně využita v analýze dat.

Pokud tedy uživatel stiskne tlačítko pro zápis do FASTA souboru, otevře se mu nové okno, které ho vyzve k výběru cílové složky uložení a zadání názvu nového souboru ve formátu FASTA. Následuje automatické vytvoření požadované struktury (z dat uvedených v tabulce) a její zápis pomocí funkce `fastawrite` Matlabu do příslušné složky.

Export výpisů do souboru aplikace Excel

Dále aplikace umožňuje export výpisů do souborů s koncovkou *„ *.xls“* programu MS Office Excel. Opět platí, že při stisknutí tlačítka *„Zapiš do excelu“* aplikace bere data, která jsou uvedena v tabulce. Následně do cell aray s daty zařadí i informaci o typu výpisu a obsahu jednotlivých sloupců.

V případě úspěšného exportu dat pomocí obou výše uvedených způsobů, uživatel obdrží kontrolní hlášku, která mu potvrdí uložení souboru, a zobrazí i cestu k němu.

Hromadný export všech souborů ve složce do souborů ve formátu FASTA

Pro snadnější a rychlejší extrakci většího počtu souborů byla vytvořena i možnost hromadného exportu souborů. Tato možnost se zpřístupní ihned po výběru souboru nebo složky a to včetně možnosti volby typu extrahovaných dat. Při stisknutí tlačítka *„Hromadný zápis do fasta“* aplikace vyzve uživatele k vybrání cílové složky uložení dat. Dále je zjištěno, které data uživatel požaduje a ve zvolené složce se vytvoří složky s příslušnými názvy. Pro hromadný export souborů je poté použit seznam všech GenBank souborů ve výchozí složce. Z každého souboru jsou extrahována data pomocí funkce `extrakce.m` a exportují se pouze data zvolená uživatelem do příslušné složky. Proces se zopakuje u každého souboru ve výchozí složce.

Jelikož uživatel nemůže volit název takto exportovaných souborů, program tyto názvy tvoří automaticky.

```
NC_005971.gbk, Bos indicus (CDS).fasta
```

V první části je uveden původní název souboru, následuje latinský název organismu a v závorce je uveden typ extrahovaných dat.

Po dokončení extrakce je uživatel upozorněn hláškou v panelu *„Více souborů“*, kde se mu následně zobrazí i místo uložení exportovaných dat.

5. Analýza rodových a mezidruhových variabilit

Pro otestování funkčnosti popsaného programu, byly vytvořeny pomocí aplikace datasety jednotlivých kódujících úseků různých organismů. Cílová analyzovaná skupina byl podkmen Vertebrata patřící do kmene Chordata (strunatci) a říše Animalia (živočichové).

Zdrojová data byla získána z databáze GenBank (dostupné z: [12]). Stažená data obsahují 61 rodů a celkem 443 druhů a jedná se o kompletní mitochondriální genomy jednotlivých organismů.

Analyzovaná data lze rozdělit do pěti skupin podle taxonomického zařazení jednotlivých organismů [20]. Rozhodující kritérium pro rozřazení organismů byl zvolen taxon třída.

Tabulka 1 Počty organismů v jednotlivých taxonech

Třída	Počet rodů	Celkový počet druhů
obojživelníci	4	30
paprskoploutví	25	190
plazi	4	29
ptáci	6	41
savci	22	152

Jednotlivé rody byly vybrány pouze v případě, že bylo nalezeno pět a více druhů a v případě, kdy mitochondriální genom obsahoval potřebné kódující úseky.

Extrakce dat pomocí vytvořené aplikace proběhla v pořádku a k další analýze byly použity pouze extrahované FASTA soubory. Vzhledem k získaným datům byla tedy vyhodnocena rodová variabilita určitých kódujících úseků a to všech třinácti mitochondriálních genů (ATP6, ATP8, COX1, COX2, COX3, CYTB, ND1, ND2, ND3, ND4, ND4L, ND5, ND6), dále úseku označujícího počátek replikace (D-loop) a obou úseků kódujících rRNA (malé a velké podjednotky).

K ohodnocení rodové variability byla stanovena vzdálenost každé sekvence vůči ostatním. Vzdálenost mezi dvěma sekvencemi je definována jako podíl počtu pozic, ve kterých se obě sekvence liší, a délky sekvence. V případě, že je zadaných sekvencí více, nejdříve se provede vícenásobné zarovnání všech sekvencí. Z tohoto zarovnání se určí konsenzuální sekvence. Konsenzuální sekvence se vytvoří tak, že na každé pozici se určí prvek, který se na dané pozici vyskytuje nejčastěji, a sekvence z těchto nečetnějších prvků každé pozice je hledaný konsenzus. Tato sekvence se následně použije ve výpočtu vzdáleností. Výpočet vzdáleností více sekvencí se provádí následovně: každá zarovnaná sekvence se porovná s konsenzuální sekvencí, z počtu rozdílných pozic pro každou sekvenci se spočítá průměr a tento průměr rozdílných pozic se vydělí délkou konsenzuální sekvence. A protože zarovnané sekvence z více

násobného zarovnání mohou obsahovat mezery, je nutné určit, zda se budou započítávat či nikoliv. V další analýze je počítáno se substitucemi i mezerami [3].

Pro zajištění automatického výpočtu rodových variabilit byl vytvořen skript `analyza.m`. Tento skript najde ve zvolené složce složky vytvořené aplikací a v nich extrahované soubory. Po té tyto soubory postupně načte a uloží do jedné proměnné tak, aby bylo možné jednotlivé kódující úseky analyzovat a vypočítat vzdálenost. Výstupem tohoto skriptu je soubor programu Excel. Tento soubor obsahuje názvy všech organismů, délku konsenzuální sekvence daného úseku a počet rozdílných pozic každé zarovnané sekvence vůči konsenzu pro daný úsek. V posledním řádku se pak nachází název rodu a vypočítaná hodnota (z výše uvedených dat) průměrné rodové variability pro daný úsek.

Tabulka 2 Příklad získaných dat pro výpočet průměrné rodové variability

	D-loop
Délka	1099
Acipenser transmontanus	233
Acipenser dabryanus	489
Acipenser stellatus	397
Acipenser gueldenstaedtii	282
Acipenser sinensis	215
Acipenser baerii	133
Acipenser schrenckii	261
Acipenser ruthenus	152
Acipenser	0,245905369

Pro názornější reprezentaci dat byly hodnoty variability vztaženy na 100 párů bází a byly zaokrouhleny na dvě desetinná místa. Výpis všech získaných hodnot je uložen v tabulce programu Excel, která se nachází v příloze této práce.

Tabulka 3 Ukázka vypočítaných hodnot průměrných rodových variabilit (vztaženo na 100 párů bází, zaokrouhleno na dvě desetinná místa)

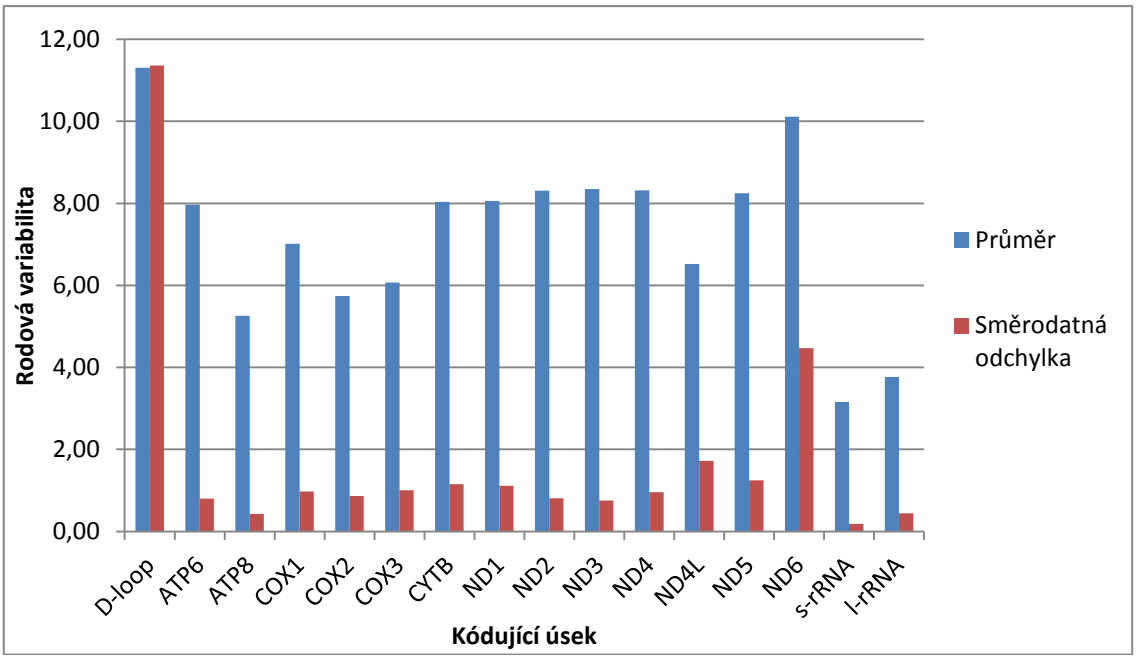
Rod	D-loop
Ambystoma	5,19
Bombina	30,97
Hynobius	4,72
Triturus	4,36

5.1. Výsledky

Pro ohodnocení vypočítaných dat byla vypočtena průměrná hodnota a směrodatná odchylka [18], z nich byl sestrojen sloupcový graf. Pro objektivní porovnání rodových variabilit byl vykreslen krabicový graf (box-plot). Ten se skládá z několika částí. V ideálním případě se uprostřed nachází hodnota mediánu, vlastní tělo („krabice“) tvoří padesát procent všech dat ohraničené dolním a horním kvartilem. Pod dolním kvartilem se nachází čtvrtina dat s nejmenšími hodnotami a nejspodnější hranice označuje minimální hodnotu. Nad horním kvartilem se naopak nachází čtvrtina nejvyšších hodnot a hranici značí maximum. V případě, že se mezi data vyskytuje příliš odlehlá hodnota nebo extrém, označí se v box-plotu hvězdičkou. Extrém je hodnota, která třikrát převyšuje kvartilový rozsah [17].

Všechny vypočítané hodnoty a získané grafy jsou přílohou této práce.

5.1.1. Třída obojživelníci

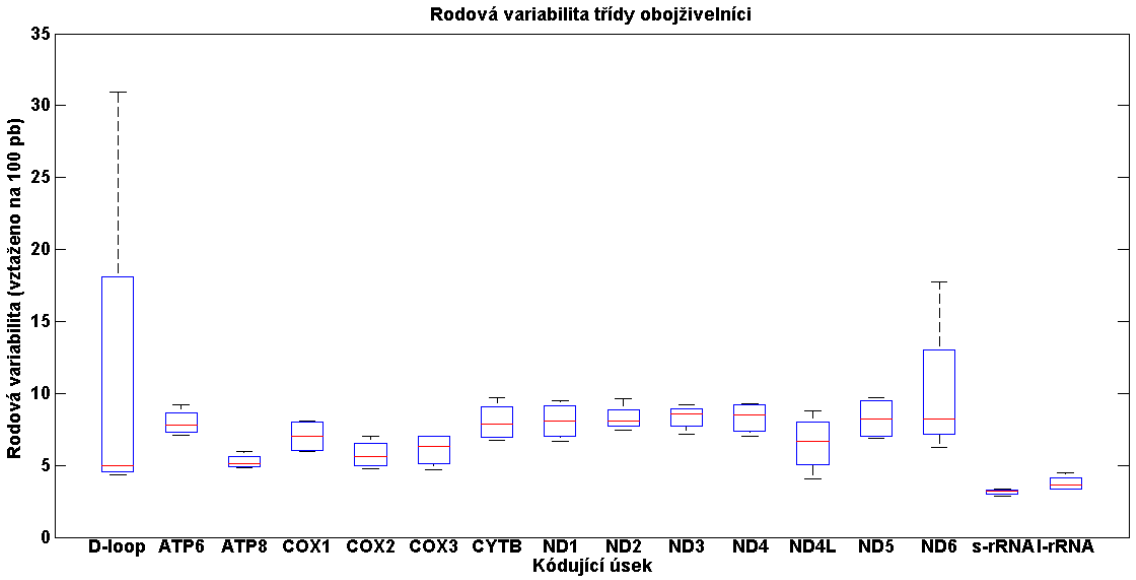


Obrázek 10 Graf průměrů a směrodatných odchylek rodových variabilit třídy obojživelníků

V třídě obojživelníci se nachází pouze čtyři rody. Průměrná hodnota rodových variabilit se výrazně liší v závislosti na kódujícím úseku. Největší průměr nalezneme u D-loop 11,31 a zde nalezneme také nejvyšší hodnotu směrodatné odchylky 11,36. Tyto vysoké hodnoty byly způsobeny vysokou rodová variabilitou rodu Bombina (česky kuňka) 30,97 oproti zbývajícím hodnotám kolem 5. Druhým nejvariabilnějším úsekem je gen ND6 s průměrnou variabilitou 10,11 a směrodatnou odchylkou 4,47. Naopak nejméně variabilní se jeví obě rRNA. rRNA menší podjednotky ribozomu s-rRNA (small rRNA)

má průměrnou variabilitu 3,16 a směrodatnou odchylku 0,18. rRNA větší podjednotky l-rRNA (large rRNA) má průměr 3,77 a směrodatnou odchylku 0,45. Nejméně variabilním genem je ATP8 s průměrem 5,26 a směrodatnou odchylkou 0,42.

Obrázek 11 Box-plot rodových variabilit třídy obojživelníků

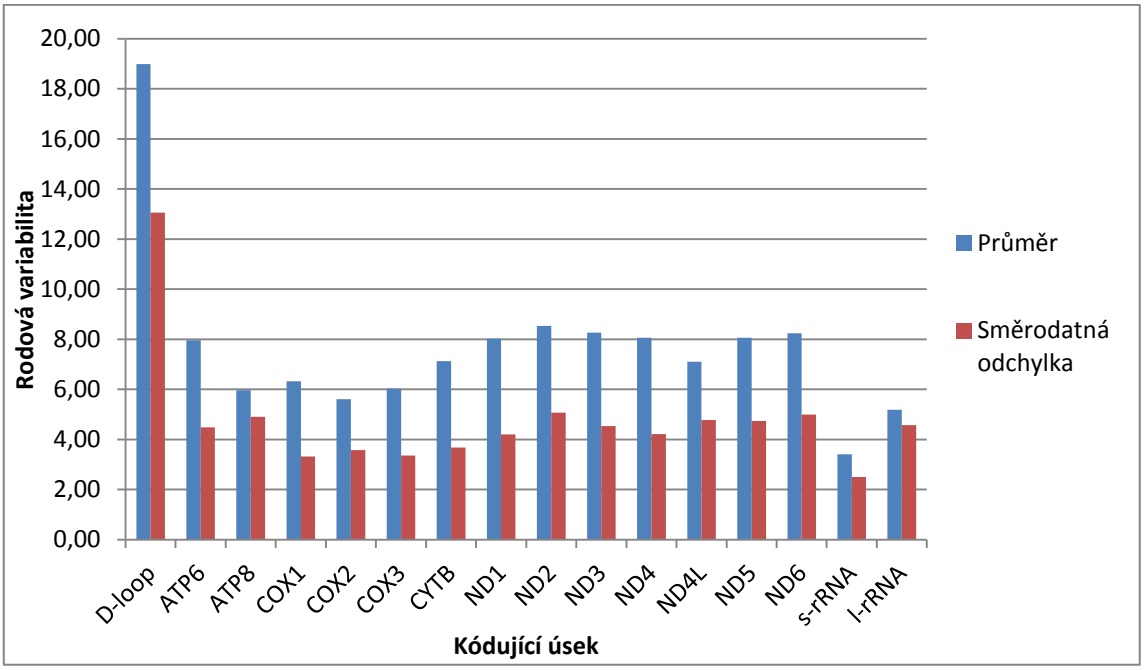


Hodnotám průměrných variabilit odpovídá i vykreslený box-plot. Z toho grafu je na první pohled patrné, které úseky jsou nejméně variabilní. Čím menší je obdélník označující rozložení hodnot průměrných rodových variabilit, tím více hodnot je soustředěno kolem mediánu. V ideálním případě úplné shody všech sekvencí bychom nalezly pouze jednu linku v hodnotě 0.

U box-plotu je opět zjevné, že nejméně variabilním úsekem je s-rRNA a nejvariabilnějším je D-loop. U D-loop je patrný výskyt ojedinělé vysoké hodnoty. Hodnota mediánu je kolem hodnoty 5 a hranice dolního kvartilu je v jeho těsné blízkosti. Zde se také projevuje nevýhoda vyhodnocování pouze na základě průměru, který může být takovými extrémním hodnotami zkreslen. Protože při bližším zkoumání jednotlivých box-plotů je patrné, že hodnota mediánu průměrných rodových variabilit u D-loop rozhodně není nevyšší. V tomto ohledu je nejvariabilnější gen ND6.

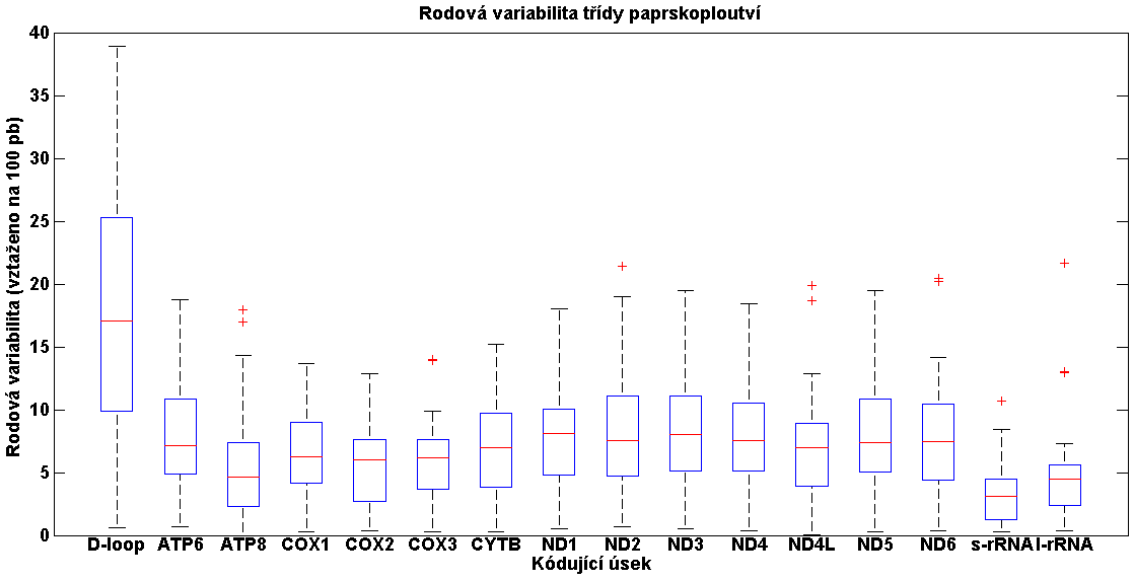
Na závěr je ovšem nutné podotknout, že vykreslené hodnoty byly počítány pouze ze čtyř hodnot průměrných rodových variabilit a pro objektivnější posouzení pouze třídy obojživelníků by bylo nutné vytvořit větší datový soubor.

5.1.2. Třída paprskoploutví



Obrázek 12 Graf průměrů a směrodatných odchylek rodových variabilit třídy paprskoploutví

Ve třídě paprskoploutví se v analýze nacházelo celkem dvacet dva rodů. Nejvyšší hodnotu průměru (spočítanou z průměrných rodových variabilit) nalezneme u úseku D-loop. Hodnota průměru je 18,99 a směrodatné odchylku 13,05 ta je také ze všech hodnot nejvyšší. Naopak nejnižší hodnotu průměru nalezneme u s-rRNA 3,14 a směrodatná odchylka je rovna hodnotě 2,50. Z analyzovaných genů je nejméně variabilním gen COX2 – průměr 5,61, směrodatné odchylka 3,58.

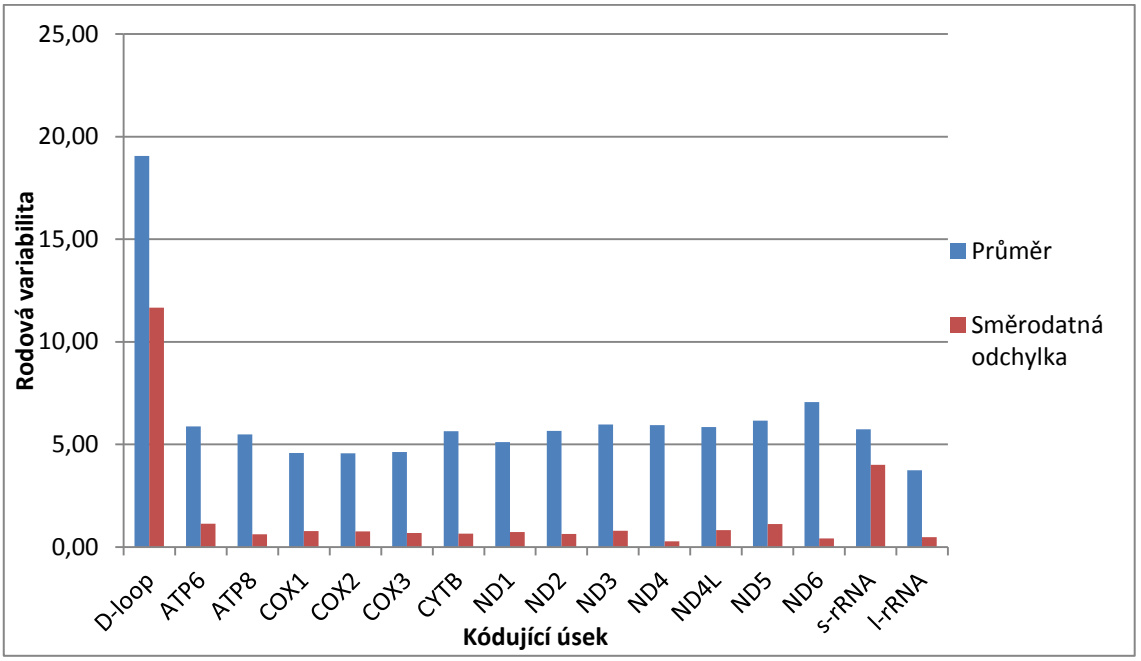


Obrázek 13 Box-plot rodových variabilit třídy paprskoploutví

U box plotu je patrné, že jeho výpočet byl použit soubor dat s velice různorodými hodnotami. Červené křížky značí polohu extrémních hodnot, které se vyskytují u většiny zkoumaných úseků a mohly významně ovlivnit výpočet průměru. Nicméně zvláště úsek s-rRNA si stále zachoval svoji nízkou variabilitu a to i u vypočítaného mediánu. D-loop je opět úsek, na kterém lze pozorovat rozložení většiny hodnot na poměrně velkém úseku 10 až 25. Extrémní hodnota tohoto úseku je 51,49 a náleží rodu Cynoglossus (česky platýs). Při určení nejméně variabilního genu z box plotu je ale situace komplikovanější. Z pohledu nejnižší hodnoty mediánu je nejméně variabilní gen ATP8, který na druhou stranu ale obsahuje dvě extrémní hodnoty též u rodu Cynoglossus a u rodu Oryzias (česky medaka). Z pohledu velikosti prostoru, který ohraničuje horní a dolní kvartil, je to gen COX3. Ten má padesát procent dat nejbližších k mediánu rozmístěných v nejmenším prostoru. Nicméně i tento úsek obsahuje dva extrémy a to opět u rodů Cynoglossus a Oryzias s hodnotami 13,95 a 13,97.

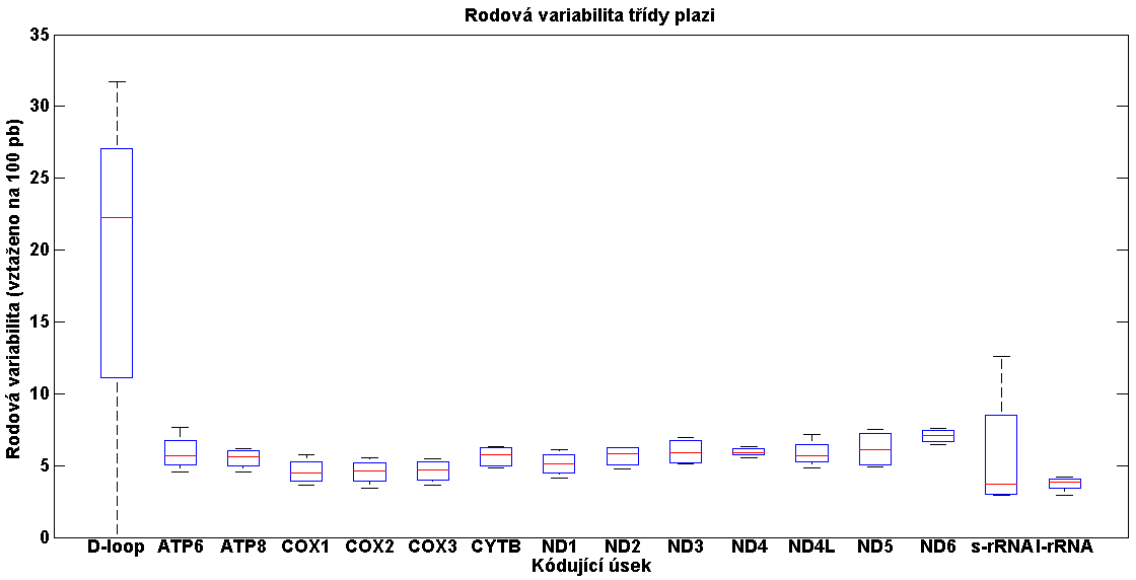
Při bližším zkoumání tabulky průměrných rodových variabilit (viz příloha) je potom patrné, že některé rody vykazují tyto extrémní hodnoty u všech zkoumaných úseků a jedná se pouze o ojedinělý extrém. Jsou to zejména rody Cynoglossus, Oryzias, Polypterus (česky bichir) a Acheilognathus (česky hořavka). Proto by bylo vhodné ověřit data, ze kterých se příslušná hodnota průměrných rodových variabilit počítala. V tomto případě, byla data vybrána správně, například se mezi nimi nenachází zástupce jiného rodu a vysoká variabilita se nachází u všech zástupců rodu. A proto byla i tato data po zvážení zařazena do analýzy.

5.1.3. Třída plazi



Obrázek 14 Graf průměrů a směrodatných odchylek rodových variabilit třídy plazi

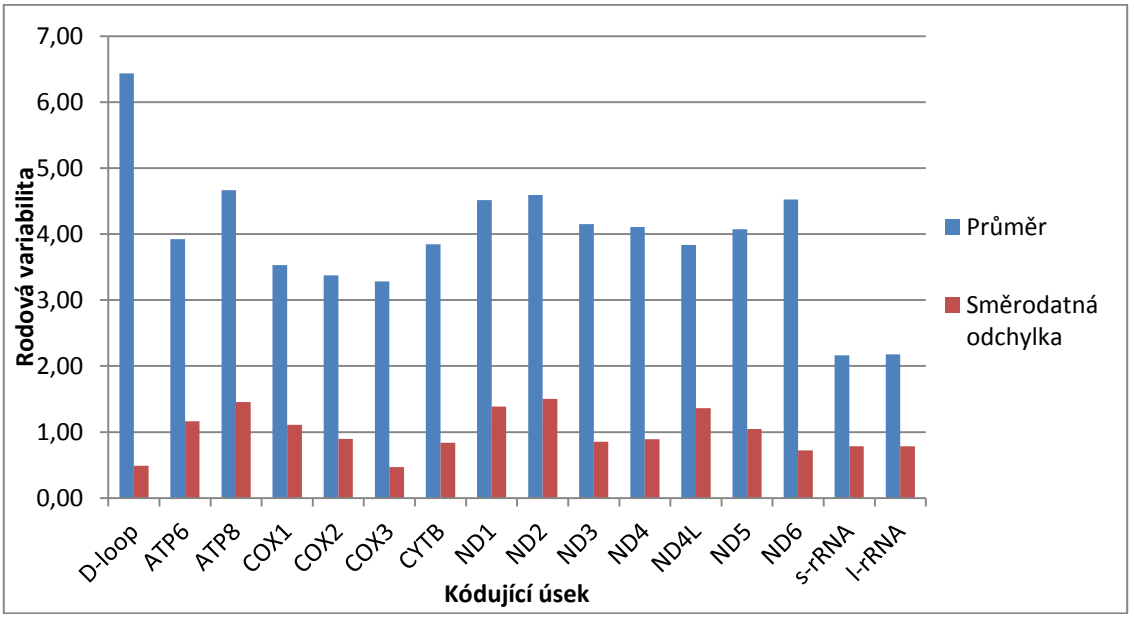
V třídě plazů byly analyzovány celkem čtyři rody. Nejvariabilnějším úsekem se opět stal úsek D-loop s průměrnou hodnotou 19,07 a směrodatnou odchylkou 11,67. Za vysokou hodnotu směrodatné odchylky zde může hlavně fakt, že jeden rod (Hypsiglena, česky užovka) měl hodnotu rodové variability 0 zatím, co ostatní rody se měly hodnoty 22 případně 32. Nejméně variabilním úsekem byl vyhodnocen úsek l-RNA průměrnou hodnotou 3,73 a směrodatnou odchylkou 0,47. Obecně lze ale pozorovat velmi nízké hodnoty směrodatných odchylek téměř u všech úseků s výjimkou D-loop a s-rRNA. U s-rRNA je to způsobeno hodnotou průměrné rodové variability u rodu Mauremys 12,61.



Obrázek 15 Box-plot rodových variabilit třídy plazi

U box-plotu lze všechny tyto anomálie zřetelně pozorovat. U D-loop lze pozorovat minimum v hodnotě 0 a vysokou variabilitu celého úseku. Stejně tak jde pozorovat i extrém u s-RNA, kde tento výkyv způsobí i nevhodné zakreslení polohy většiny dat. Tudíž by směrodatnějším ukazatelem v tomto případě měla být hodnota mediánu. Nicméně nejméně variabilním úsekem je v této analýze úsek l-rRNA. Všechny geny tohoto rodu jsou také srovnatelně variabilní. Otázkou ovšem zůstává, jak by se tyto variability změnili při použití většího počtu rodů této třídy.

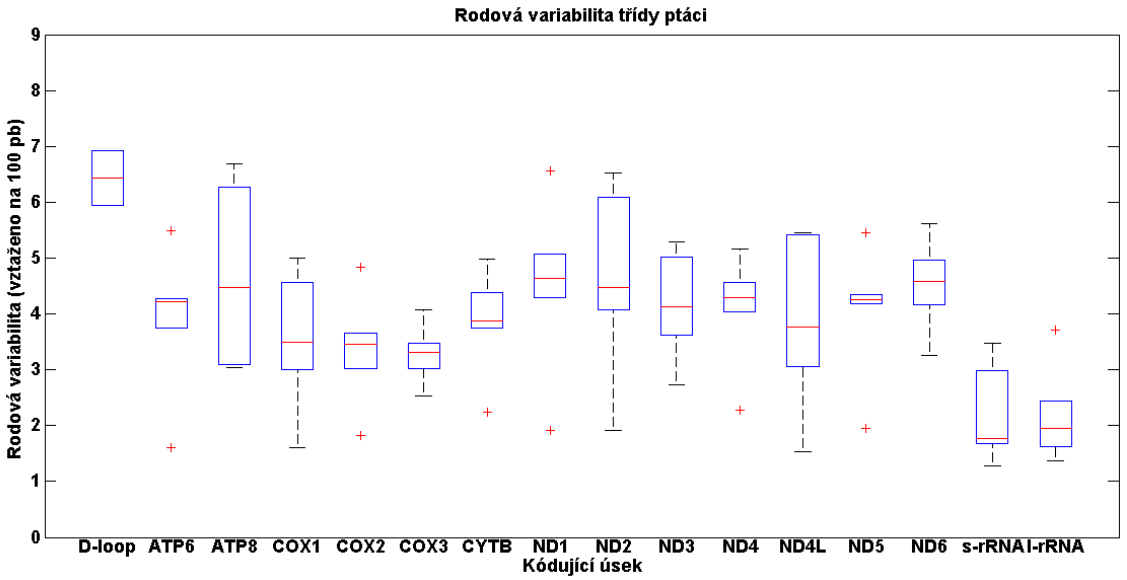
5.1.4. Třída ptáci



Obrázek 16 Graf průměrů a směrodatných odchylek rodových variabilit třídy ptáci

U třídě ptáci bylo analyzováno šest rodů. Jelikož k výpočtu průměru D-loop byly použity pouze dvě hodnoty, nemá význam tento úsek hodnotit. Hodnoty byly pouze dvě, protože u některých organismů nebyl tento úsek ve zdrojovém souboru zapsán. Nicméně zbylé úseky popsány byly a tak byly tyto organismy také použity v analýze. V případě vyřazení těchto organismů by se použité datasety značně zredukovaly.

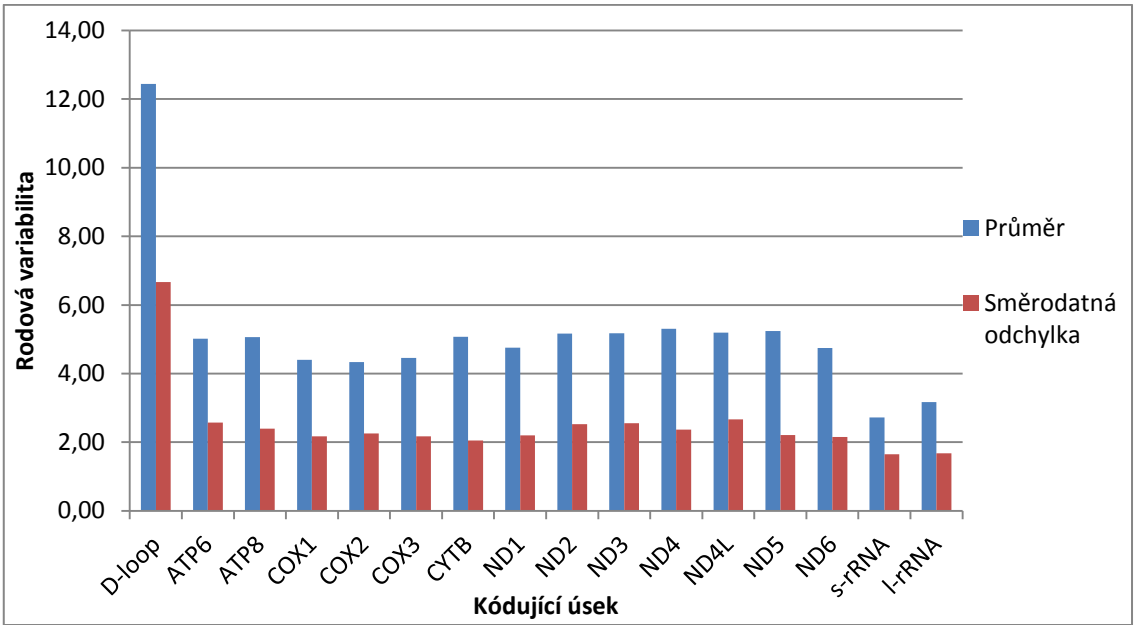
Z pohledu průměrů a směrodatných odchylek byl nejvariabilnější gen ND2 s hodnotou průměru 4,59 a směrodatnou odchylkou 1,50. Nejméně variabilní je úsek s-rRNA s průměrem 2,16 a směrodatnou odchylkou 0,79.



Obrázek 17 Box-plot rodových variabilit třídy ptáci

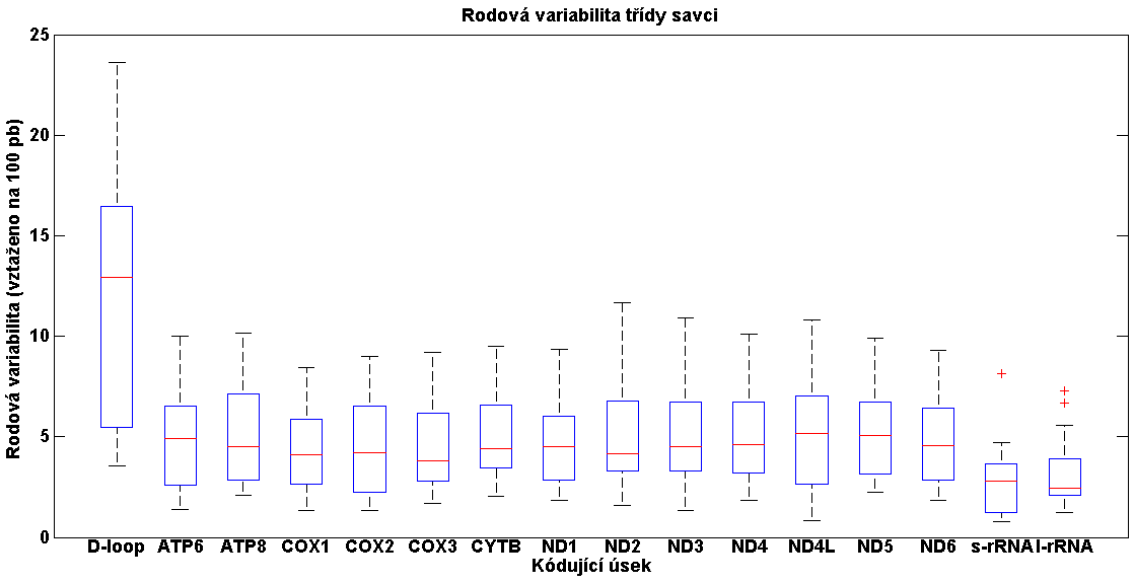
U box-plotu lze pozorovat značné množství extrémů: maxima i minima. Značná část kódujících úseků je velmi variabilní. Především geny ATP8,COX1, ND2 či ND4L. Při podrobném zkoumání hodnot lze zjistit, že extrémní hodnoty v podobě minim způsobují hodnoty rodu Gallus (česky kur). Za maximy lze nalézt hodnoty rodu Tachycineta (česky vlaštovka). A velké rozpětí těchto hodnot při tak malém počtu hodnot zapříčiňuje nemožnost objektivního zhodnocení. I za těchto podmínek ovšem najdeme úsek, který těmito extrémy není výrazně ovlivněn. Jsou to gen COX3. Případně úsek s-rRNA.

5.1.5. Třída savci



Obrázek 18 Graf průměrů a směrodatných odchylek rodových variabilit třídy savci

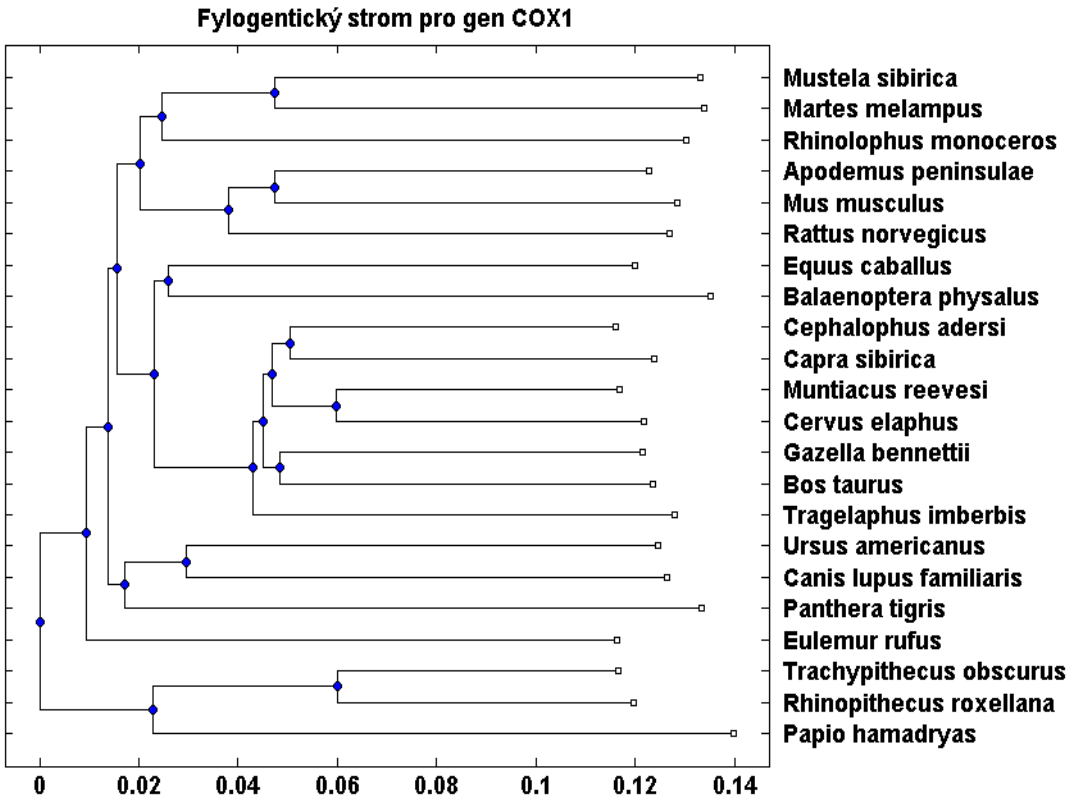
Z průměrných rodových variabilit dvaceti dvou rodů byly vypočítány průměrné hodnoty a směrodatné odchylky. Nejvariabilnější úsekem je opět D-loop s hodnotami průměru 12,44 a směrodatné odchylky 6,67. Variability všech genů jsou poměrně vyrovnané, nicméně nejméně variabilní úseky jsou geny COX1 s průměrem 4,40 a odchylkou 2,18 a gen COX2 s průměrem 4,34 a odchylkou 2,25. Ze všech úseků má ovšem nejmenší průměr a směrodatnou odchylku úsek s-rRNA s hodnotami 2,73 a 1,65.



Obrázek 19 Box-plot rodových variabilit třídy savci

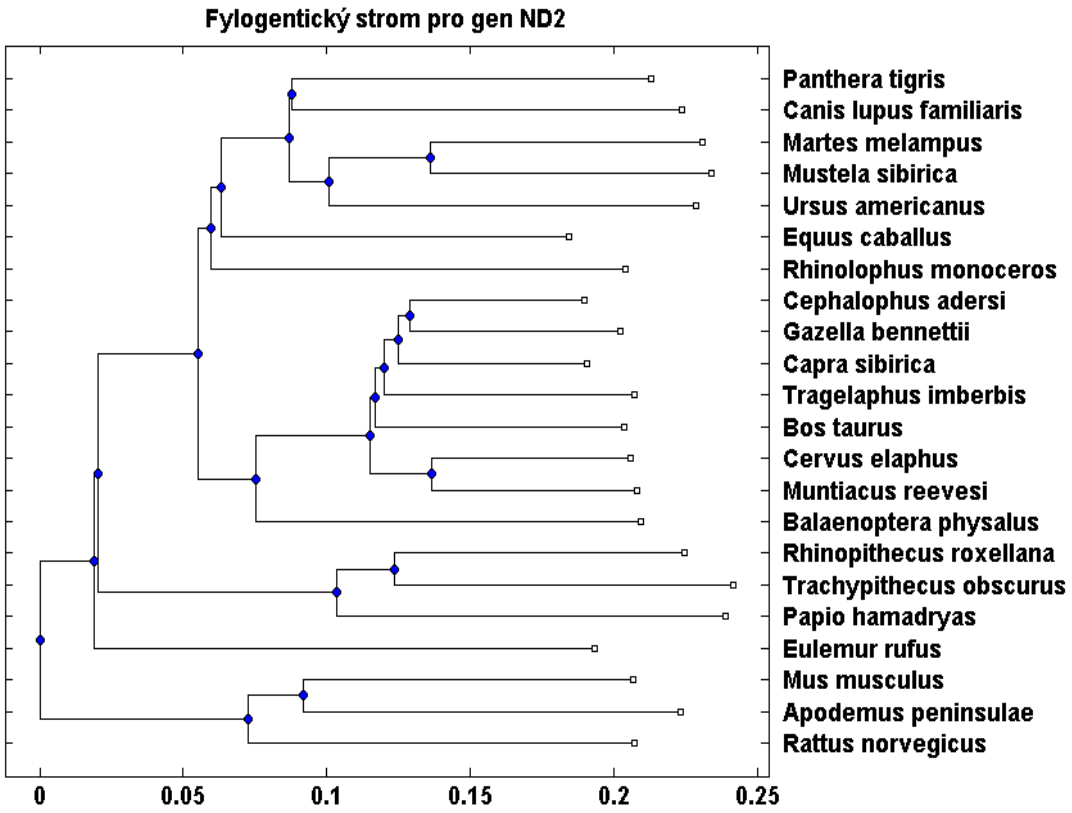
U krabicového grafu se tyto závěry potvrzují. Nejvariabilnější je úsek D-loop. Nejméně variabilní je s-rRNA, i když zde lze nalézt jeden extrém u rodu Capra (česky koza) s hodnotou 8,15. U l-rRNA jsou to dokonce dva extrémy – hodnoty rodu Apodemus (česky myšice) 6,69 a Trachypithecus (česky hulman) 7,27. Nejméně variabilní gen podle box plotu je gen COX1, ale nutné podotknout, že podobný průběh box-plotu má i gen CYTB či ND1.

Dále byly u třídy savci vytvořeny fylogenetické stromy. Z každého rodu byl vybrán jeden druh, takže výsledné stromy budou mít 22 listů. Stromy byly vykresleny pouze pro tři úseky: pro gen COX1 (nejméně variabilní u savců), gen ND2 (nejvíce variabilní u savců) a s-rRNA (nejméně variabilní kódující úsek). A hodnotícím kritériem byla stanovena věrnost vykresleného fylogenetického stromu vůči známé taxonomii [19], [20]. Pro výpočet vzdáleností mezi jednotlivými druhy byla zvolena metoda Junkes-Cantor a pro zrekonstruování stromů metoda neighbor-joining [3].



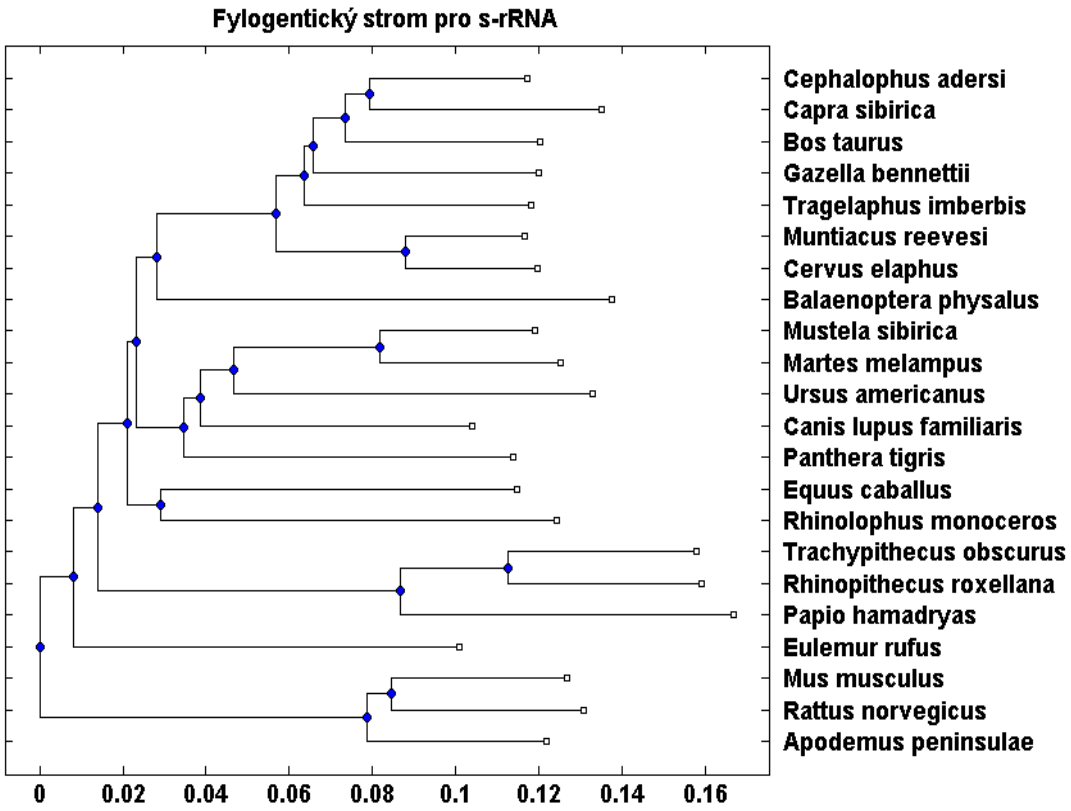
Obrázek 20 Fylogenetický strom pro gen COX1

U fylogenetického stromu pro gen COX1 je patrné, že neodpovídá teoretickým předpokladům. Zejména tento strom zařazuje druhy řádu šelem na různé větve stromu. Větev s druhy *Mustela sibirica* (česky lasice sibiřská), *Martes melampus* (česky sobol východní) by měla být daleko blíže větvi druhů *Ursus americanus* (česky medvěd baribal), *Canis lupus familiaris* (česky pes domácí) a *Panthera tigris* (česky tygr). Dále je zde špatně přiřazen *Eulemur rufus* (česky lemur červenavý), který není napojen přímo na větev primátů. Zbylé organismy jsou již zařazeny správně.



Obrázek 21 Fylogenetický strom pro gen ND2

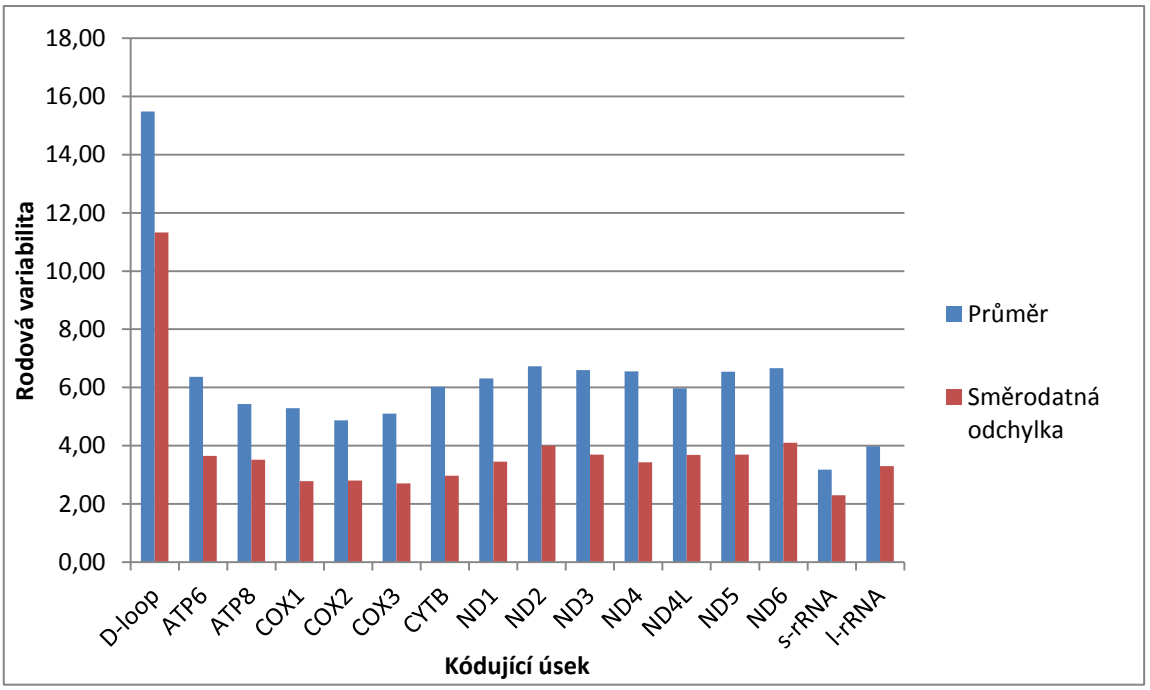
U nejvariabilnějšího genu ND2 též lze nalézt řadu nesrovnalostí. Zaprvé strom přiřazuje druh *Equus caballus* (česky kůň) řádu lichokopytníků k řádu šelem. Dále špatně zařazuje druh *Eulemur rufus*, který není přiřazen blíže k primátům, ale rozdíl vzdáleností, který rozhodl o špatném zařazení je minimální. Jinak fylogenetický strom zhruba odpovídá teoretickým předpokladům, správně spojuje větve řádu sudokopytníků i hlodavců.



Obrázek 22 Fylogenetický strom pro s-rRNA

Poslední strom vytvořený na základě úseku s-rRNA pro různé druhy savců odpovídá nejvíce reálnému základu. Všechny živočichy správně zařazuje do společných větví. Jen *Eulemur rufus* není přímo napojen na větve primátů. Vzhledem k nízké variabilitě tohoto úseku, která se potvrdila v předchozí analýze, a relativní správnosti vytvořeného fylogenetického stromu, vykazuje tento úsek velmi zajímavé vlastnosti, které by mohly být použity při taxonomickém zařazení neznámého organismu.

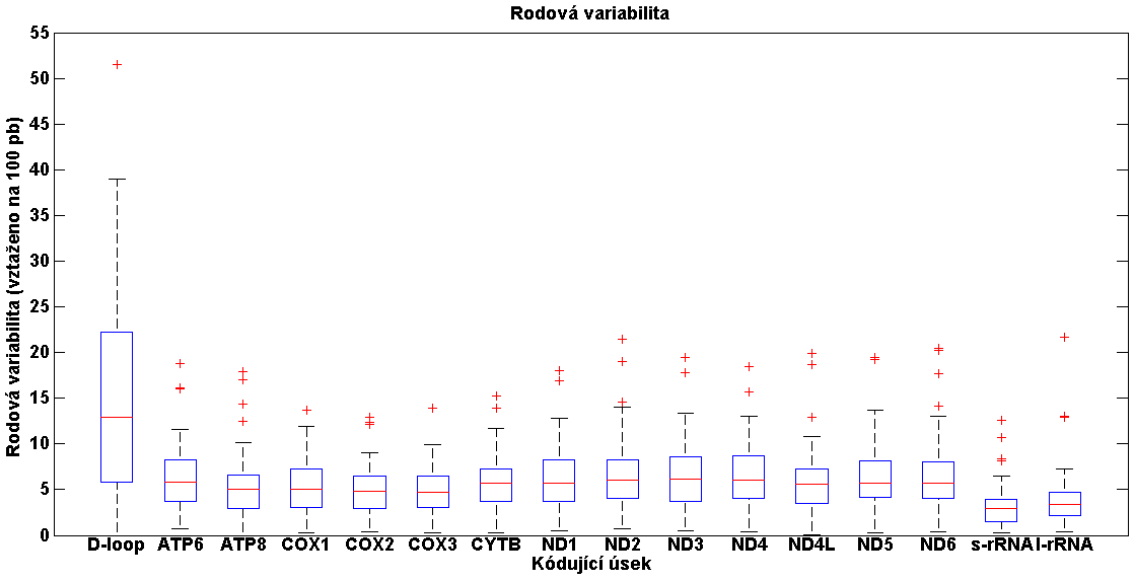
5.1.6. Souhrn



Obrázek 23 Graf průměrů a směrodatných odchylek rodových variabilit

Pro objektivní posouzení průměrné rodové variability jednotlivých úseků je nutné do analýzy zahrnout všechny rody každé třídy. Jak bylo popsáno v předešlých odstavcích, čím více hodnot bude použito pro analýzu, tím více realitě odpovídající výsledek bude získán a budou moci být vyvozeny odpovídající závěry.

Z pohledu vypočítaného průměru rodových variabilit, lze jednoznačně říci, že úsek D-loop je nejvíce variabilní. Stejně jak tomu bylo u analýz jednotlivých tříd organismů. Průměrná hodnota D-loop pro celý dataset je 15,48, směrodatná odchylka má hodnotu 11,32 a je též nejvyšší hodnotou. Nejnižší průměr 3,18 a nejnižší směrodatnou odchylku 2,30 lze nalézt u úseku s-rRNA. V rámci analyzovaných genů je nejméně variabilní gen COX2 (s průměrem 4,87 a směrodatnou odchylkou 2,80), po té gen COX3 (5,10 a 2,71). Naopak nejvariabilnějším genem je ND2 (6,73 a 4,00).



Obrázek 24 Box-plot rodových variabilit

Z box-plotu je patrné, že větší množství dat umožnilo správně identifikovat extrémy a tak lépe objektivně analyzovat získaná data. Potvrdilo se tedy, že D-loop je nejvíce variabilní. U s-rRNA je zajímavé pozorovat značný výskyt extrémů, ale i tak je tento úsek nadále nejméně variabilní. Z genů jsou to COX2 a COX3, které jsou nejméně variabilní. Jejich opakem je gen ND2 či ND6.

6. Závěr

Prvním bodem bakalářské práce bylo seznámení se s databází GenBank a používanými datovými formáty (zejména s datovým formátem *.gbk). Genbank patří do skupiny veřejně přístupných databází nukleových sekvencí. Spolu s EMBL a DDBJ tvoří skupinu provázaných databází, které mají stejný obsah a každý den se synchronizují. Nejpoužívanějším datovým formátem je GenBank, případně FASTA. GenBank se skládá ze tří částí: hlavičky, anotace a sekvence. Anotace je rozdělena do mnoha polí: název, identifikační údaje, délka sekvence, typ molekuly a její topologie, definice, fylogenetické zařazení, reference, údaje o kódujících úsecích sekvence a mnohé další. Datový formát FASTA tyto informace redukuje na hlavičku a sekvenci.

Kromě jaderné DNA můžeme v buňce nalézt i DNA mimojadernou mitochondriální (mtDNA). Mitochondrie jsou semiautonomní buněčné organely, které vznikly endosymbiózou, a pro buňku syntetizují molekuly ATP za současného spotřebovávání kyslíku. Mitochondrie se skládají ze dvou membrán a obsah vnitřní membrány se nazývá matrix. Zde najdeme mimo jiné i mtDNA. Mitochondriální DNA má díky svému původu kruhovitý tvar stejně jako prokaryotická DNA. mtDNA má u člověka délku 16 569 pb a kóduje 2 molekuly rRNA, 22 molekul tRNA a 13 proteinů pro oxidační fosforylaci, celkem tedy zde najdeme 37 kódujících úseků. U rostlin kromě mitochondrií může nalézt i další druh semiautonomních organel plastidy. Plastidů rozlišujeme širokou škálu od nejjednoduššího proplastu, ze kterého se vyvíjí ostatní typy, až po nejznámější chloroplast, ve kterém probíhá fotosyntéza. Tyto organely též obsahují DNA – plastidovou někdy též nazývanou chloroplastovou (cp DNA), která má tvar složitější. Od dosud předpokládaného tvaru kruhového se upouští a byla prokázána přítomnost lineárních molekul cpDNA. V dnešní době bylo sekvenováno mnoho mitochondriálních a chloroplastových genomů, nicméně význam všech kódujících i nekódujících úseků sekvencí zatím nebyl objasněn a je předmětem dalších studií a výzkumů.

Pro zpracování GenBank souborů v Matlabu existuje bioinformatický toolbox. Pro jejich načtení slouží funkce `genbankread`, jejímž výstupem je struktura obsahující stejná data jako původní soubor. Funkce provádí načítání celých souborů a to podle přesných klíčových slov, tudíž je jednotnost zápisu dat v genbank formátu nutností. Pro vyhodnocení úspěšnosti funkce bylo načteno 3 873 souborů. U těchto souborů byla zkontrolována přítomnost sekvence a CDS. Sekvence byly načteny u všech souborů správně. V případě CDS některé soubory tuto položku neobsahovaly a to z důvodu absence CDS již v originálním souboru. Na druhou stranu nepřítomnost CDS neznamená, že daná sekvence žádný gen nekóduje. Pouze vypovídá o tom, že autor

příspěvku tuto položku z nějakého důvodu nevyplnil. Tudíž je otázkou, na kolik správné jsou příspěvky v databázi a na kolik se dodržují pravidla zápisu.

A proto byla vytvořena nová uživatelská aplikace, která `genbankread` do jisté míry nahrazuje a umožňuje uživateli další práci s GenBank souborem. Primárně byla aplikace vyvinuta pro práci se soubory mitochondriálních a plastidových genomů. Zejména nabízí uživateli stručný přehled informací o daném souboru. Vypisuje seznam jednotlivých kódujících úseků (D-loop, CDS, rRNA a tRNA). Poskytuje přehled dalších statistik například výpis úseků, které nebyly zařazeny ani do jedné z výše uvedených kategorií. Tyto úseky ovšem svůj význam mají, protože v mitochondriální ani plastidové DNA by se žádné nekódující úseky neměly vyskytovat. Dále aplikace umožňuje exportovat získaná data a to data jednoho souboru nebo všech souborů ve vybrané složce. Volba požadovaných dat je ponechána na uživateli a extrahované úseky jsou uloženy ve formátu FASTA.

Následně byla tato aplikace použita pro vytvoření datasetů mitochondriálních sekvencí skupiny Vertebrata. Celkem bylo použito 443 druhů organismů, které lze rozčlenit do 61 rodů. Pro tyto rody byla následně provedena analýza průměrných rodových variabilit a u zástupců třídy savců bylo provedeno vykreslení fylogenetických stromů na základě mezidruhové variability (pouze u genů COX1, ND2 a u úseku s-rRNA). Cílem analýzy bylo určení variability všech třinácti genů, úseku D-loop a obou rRNA. Z analýz bylo zjištěno, že nejméně variabilním úsekem je s-rRNA. Fylogenetický strom vykreslený na základě mezidruhové variability tohoto úseku také nejvíce odpovídal známé taxonomii. Hodnoty průměrných rodových variabilit všech genů jsou velmi podobné a o žádném genu nelze jednoznačně říci, že je nejméně variabilní. Nízkých hodnot dosahují geny COX1, COX2 či COX3. Obecně nejvariabilnějším úsekem je jednoznačně D-loop.

V neposlední řadě je nutné podotknout, že uvedená analýza byla provedena na relativně malém souboru organismů a soužila pouze k ověření funkčnosti aplikace na extrakci dat. Pro dosažení směrodatných výsledků analýzy, aby měly být výpočty provedeny na větším datovém souboru a pomocí více metod hodnocení variability. V případě vypracování takovéto práce, by se mohla vytvořená aplikace stát vhodným nástrojem pro práci se vstupními daty.

7. Seznam použité literatury

[1] SCHEFFLER, I. E. Mitochondria. 2nd ed., Wiley-Blackwell, 2007, 472 s. ISBN 978-0-470-04073-7.

[2] BENDICH, A. Circular Chloroplast Chromosomes: The Grand Illusion. The Plant Cell. 2004, 16, 1661-1666.

[3] BAXEVANIS, Andreas D. a B. F. Francis OUELLETTE. Bioinformatics: A Practical Guide To The Analysis Of Genes And Proteins. 2nd ed. New York: Wiley-Interscience, 2001. ISBN 978-047-1223-924.

[4] GenBank Celebrates 25 Years of Service with Two-Day Conference: Leading Scientists Will Discuss the DNA Database at April 7-8 Meeting. The National Institues of Health [online]. 2008 [cit. 2015-01-02]. Dostupné z: http://www.nih.gov/news/health/apr2008/nlm-03.htm

[5] STRASSER, Bruno. Genetics: GenBank: Natural History in the 21st Century? Science. 2008, č. 322.

[6] GenBank Flat File Release 205.0: Distribution Release Notes. The National Center for Biotechnology Information [online]. 2014 [cit. 2014-12-15]. Dostupné z: ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt

[7] GenBank Sample Record. The National Center for Biotechnology Information [online]. 2006 [cit. 2014-12-15]. Dostupné z: http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html#LocusA

[8] ALBERTS, Bruce, et al. Základy buněčné biologie: Úvod do molekulární bilogie. 2. vyd. Translation Prof. RNDr. Arnošt Kotyk, DrSc. Ústí nad Labem: Espero publishing, 1998. ISBN 80-902906-2-0.

[9] PIERCE, B. A. Genetics – A Conceptual Approach. 4th ed, W. H. Freeman, ISBN: 9781464109461

[10] HUDÁK, Ján. Chloroplasty: Zelené organely. *ŽIVA*. 2010, č. 3.

[11] Bioinformatics Toolbox Documentation. MathWorks: MATLAB and Simulink for Technical Computing [online]. 2014 [cit. 2015-01-02]. Dostupné z: http://www.mathworks.com/help/bioinfo/index.html

[12] National Center for Biotechnology Information [online]. 2015 [cit. 2015-01-02]. Dostupné z: http://www.ncbi.nlm.nih.gov/

[13] VŠCHT Praha, Ústav organické technologie. Mitochondriální dědičnost. [online]. [cit. 2015-01-02]. Dostupné z: http://www.vscht.cz/kot/resources/studijnimaterialy/bc-skripta/kapitola04.pdf

[14] Rostlinná cytologie. Katedra experimentální biologie rostlin PřF UK. [online]. [cit. 2015-01-02]. Dostupné z : http://kfrserver.natur.cuni.cz/lide/schwarze/cytologie/mitochondrie%2013.pdf

[15] ŠKALOUD, Pavel. Plastidy a jejich původ – přednáška. Protistologie [online] [cit. 2015-01-02]. Dostupné z : http://www.protistologie.cz/

[16] Mitochondrial DNA. Genetics Home Reference [online] [cit. 2015-01-02]. Dostupné z: http://ghr.nlm.nih.gov/mitochondrial-dna

[17] Krabicový graf (boxplot). Eistat. [online] [cit. 2015-05-26]. Dostupné z: http://www.eistat.cz/popis/boxplot/index.htm

[18] Statistika: Základní pojmy. Gymnázium Elgartova [online] [cit. 2015-05-26]. Dostupné z: http://www.gymelg.cz/sites/default/files/chemie/Statistika_Studenti.pdf

[19] Common Taxonomy Tree. Taxonomy Browser. [online] [cit. 2015-05-26]. Dostupné z: http://www.ncbi.nlm.nih.gov/Taxonomy/CommonTree

[20] BioLib: Taxonimic tree of plants and animals with photos. [online] [cit. 2015-05-26]. Dostupné z: http://www.biolib.cz/cz/main/

8. Seznam zkratek

gb, gbk	GenBank
NCBI	National Center for Biotechnology Information
NIH	National Institutes of Health
DDBJ	The DNA Data Bank of Japan
EMBL	the European Molecular Biology Laboratory
EBI	the European Bioinformatics Institute
RefSeq	NCBI Reference Sequence Database
INSDC	the International Nucleotide Sequence Database Collaboration
ORF	otevřený čtecí rámec
kpb	kilo párů bází
bp	párů bází
DNA	deoxyribonukleová kyselina
mtDNA	mitochondriální deoxyribonukleová kyselina
cpDNA	chloroplastová deoxyribonukleová kyselina
RNA	ribonukleová kyselina
tRNA	transferová ribonukleová kyselina
rRNA	ribozomální ribonukleová kyselina
s-rRNA	ribozomální ribonukleová kyselina malé („small“) podjednotky ribozomu
l-rRNA	ribozomální ribonukleová kyselina velké („large“) podjednotky ribozomu
NAD1	gen NADH dehydrogenázy podjednotky 1
NAD2	gen NADH dehydrogenázy podjednotky 2
NAD3	gen NADH dehydrogenázy podjednotky 3
NAD4	gen NADH dehydrogenázy podjednotky 4
NAD4L	gen NADH dehydrogenázy podjednotky 4L
NAD5	gen NADH dehydrogenázy podjednotky 5
NAD6	gen NADH dehydrogenázy podjednotky 6
ATP8	gen ATP syntázy podjednotky 8
ATP6	gen ATP syntázy podjednotky 6
COX1	gen cytochromoxidázy podjednotky 1
COX2	gen cytochromoxidázy podjednotky 2
COX3	gen cytochromoxidázy podjednotky 3
CYTB	gen cytochrom b

9. Seznam příloh

Příloha 1: CD s digitální verzí diplomové práce, zdrojovými soubory programu, obrázkem uživatelského rozhraní vytvořené aplikace, zdrojovými FASTA soubory použitých sekvencí, tabulkami rodových variabilit, obrázky fylogenetických stromů a boxplotů